Virology 546 (2020) 51–66

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

New insights into the evolutionary features of viral overlapping genes by discriminant analysis

Angelo Pavesi

Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area Delle Scienze 23/A, I-43124, Parma, Italy

ARTICLE INFO

Keywords:
Ancestral reading frame
Apoptin
Codon usage
Computer simulation
De novo reading Frame
Homologs
Multivariate statistics
SARS-CoV-2
Therapeutic proteins
Virus evolution

ABSTRACT

Overlapping genes originate by a mechanism of overprinting, in which nucleotide substitutions in a pre-existing frame induce the expression of a *de novo* protein from an alternative frame. In this study, I assembled a dataset of 319 viral overlapping genes, which included 82 overlaps whose expression is experimentally known and the respective 237 homologs. Principal component analysis revealed that overlapping genes have a common pattern of nucleotide and amino acid composition. Discriminant analysis separated overlapping from non-overlapping genes with an accuracy of 97%. When applied to overlapping genes with known genealogy, it separated ancestral from *de novo* frames with an accuracy close to 100%. This high discriminant power was crucial to computationally design variants of *de novo* viral proteins known to possess selective anticancer toxicity (apoptin) or protection against neurodegeneration (X protein), as well as to detect two new potential overlapping genes in the genome of the new coronavirus SARS-CoV-2.

1. Introduction

The viral genome often contains more coding information than a co-linear relationship between nucleotide and protein sequences would suggest. This increase in coding density depends on overlapping genes, in which two different reading frames on the same strand are translated to yield two proteins (Normark, 1983). Overlapping genes originate by a mechanism of overprinting, in which critical nucleotide substitutions in a pre-existing ancestral frame induce the expression of a *de novo* protein from an alternative frame (Keese and Gibbs, 1992; Rancurel et al., 2009; Sabath et al., 2012). The *de novo* proteins, unlike the ancestral ones, usually lack any remote homologs in protein databases (Gibbs and Keese, 1995).

Overlapping genes were first detected in the genome of the bacteriophage ΦX174 by Barrell and Air (1976). For many years they were thought to be limited to viruses, but experimental (Michel et al., 2012; Bergeron et al., 2013; Vanderperre et al., 2013; Fellner et al., 2015) and computational studies (Chung et al., 2007; Delaye et al., 2008; Ribrioux et al., 2008; Vanderperre et al., 2012) indicate that they also occur in prokaryotes and eukaryotes.

Thus, the expression of two proteins from the same mRNA has changed the traditional view that a mature eukaryotic mRNA is a monocistronic molecule with a single translated open reading frame (ORF) (reviewed by Mouilleron et al., 2016; Brunet et al., 2018). Interestingly, it has been found that some human cancer-specific antigens, silent in normal tissues, are translated from alternative open reading frames (AltORFs) (Wang et al., 1996, 1998; Rosenberg et al., 2002; Mandic et al., 2003; Slager et al., 2003). These neoantigens are promising targets for the development of antitumour immunotherapies with a potentially broader coverage of patients (Smith et al., 2019).

Researchers have advanced two theories, not mutually exclusive, to explain the abundance of overlapping genes in viruses. The gene-compression theory claims that this feature is a strategy to maximize the information content of small genomes (Miyata and Yasunaga, 1978; Lamb and Orvath, 1991; Pavesi et al., 1997; Krakauer, 2000; Peleg et al., 2004), as a consequence of error-prone polymerases (Belshaw et al., 2007) and biophysical constraints acting on the capsid structure (Chirico et al., 2010). The gene-novelty theory asserts that the birth of *de novo* proteins, often overprinting a structural or enzymatic protein, is driven by selection pressures providing the virus with a fitness advantage that lead to their fixation (Keese and Gibbs, 1992; Rancurel et al., 2009; Brandes and Linial, 2016).

Indeed, the *de novo* proteins originated by overprinting often play a critical role in virus infection, by neutralizing the host interferon response (van Knippenberg et al., 2010; McFadden et al., 2011; Wensman et al., 2013) and the RNA interference pathway (Vargason et al., 2003; Chellappan et al., 2005; Li and Ding, 2006), by inducing apoptosis in host cells (Noteborn et al., 1994; Chen et al., 2001; Boehme et al., 2013) or promoting the systemic spread of virus (Taliansky et al., 2003), A few *de novo* proteins can also exert functions that are not virus-specific.

E-mail address: angelo.pavesi@unipr.it.

https://doi.org/10.1016/j.virol.2020.03.007
Received 20 March 2020; Accepted 29 March 2020
Available online 02 April 2020
0042-6822/ © 2020 Elsevier Inc. All rights reserved.

A well-known example is the apoptin of *Chicken anemia virus*, which induces cell death in a broad range of human tumour cell lines but not in normal cells (Danen-Van Oorschot et al., 1997; Leliveld et al., 2003; Backendorf et al., 2008; reviewed by Castro et al., 2018). Another example is the X protein of *Borna disease virus*, which shows protective properties against neurodegeneration *in vitro* and *in vivo* (Szelechowski et al., 2014). To increase their therapeutic effectiveness, researchers have constructed N-terminal deleted forms or short peptides both for apoptin (Shen Ni et al., 2013; Ruiz-Martínez et al., 2017; Zhang et al., 2017) and X protein (Szelechowski et al., 2014; Ferré et al., 2016).

To identify viral overlapping genes by sequence analysis, Firth (2014) and Sealfon et al. (2015) have developed statistical methods that detect, in a set of homologous protein-coding regions, the atypical pattern of nucleotide substitution induced by the overlap (i.e. a significantly reduced rate of synonymous substitution). Schlub et al. (2018) have recently proposed a new statistical method. It examines single gene or genome viral sequences, and can therefore be applied in many situations where the previous methods could be ineffective.

To provide a benchmark for systematic studies, we assembled a dataset of 80 overlapping genes 180 nt or longer from eukaryotic viruses with a genome shorter than 30 kb, whose expression is supported by reliable experimental evidence (S1 dataset from Pavesi et al., 2018). A first analysis revealed that overlapping genes differ significantly from non-overlapping genes in their nucleotide and amino acid composition (Pavesi et al., 2018). A further analysis, carried out on 75 pairs of homologous overlaps, revealed that half of overlapping genes undergo asymmetric evolution, as the protein encoded by one frame is significantly more variable than that encoded by the other frame (Pavesi, 2019).

In the present study, I enlarged the dataset of overlapping genes by increasing the number of homologs. The dataset I assembled contains 319 viral overlapping genes, instead of 80 or 150 as in the previous studies (Pavesi et al., 2018, Pavesi, 2019). I first investigated the nucleotide and amino acid composition of overlapping genes with the principal component analysis (PCA) (Hotelling, 1936), by using as a control the entire complement of non-overlapping genes in the viral genome.

I then analyzed the dataset with the partial least squares-discriminant analysis (PLS-DA) (reviewed by Lee et al., 2018) and the Fisher's linear discriminant analysis (LDA) (Fisher, 1936), with the aim to separate overlapping from non-overlapping genes with the best accuracy. To test which of the two theories about the abundance of gene overlap in viruses (gene-compression and gene-novelty) is the most plausible one, I also applied LDA to a subset of overlapping genes with known genealogy (which frame is ancestral and which one is *de novo*).

Finally, I developed a computational algorithm that simulated the birth of new overlapping genes encoding variants of two peculiar proteins: the apoptin of *Chicken anemia virus* (apoptin is the *de novo* protein overprinting the ancestral capsid protein VP2) and the X protein of *Borna disease virus* (X is the *de novo* protein overprinting the ancestral phosphoprotein). The huge amount of new overlapping genes yielded by simulation was processed using PLS-DA and LDA, with the aim to obtain a set of protein variants whose selective anticancer activity (apoptin) or protection against neurodegeneration (X protein) could be tested experimentally.

2. Materials and methods

2.1. Search for homologous overlapping genes in the NCBI database

I first added to the dataset of 80 overlapping genes (S1 Dataset from Pavesi et al., 2018) two other overlaps experimentally proven: NS1 protein/NS2 protein of *Influenza A virus* (Lamb, 1980) and large T antigen/ALTO protein of *Merkel cell polyomavirus* (Carter et al., 2013). I then extracted from the dataset the amino acid sequence of the two proteins encoded by each overlap. For each protein, I searched for homologs against the non-redundant protein sequences and the reference protein sequences NCBI database using BLASTP with an *E* cutoff of 10^{-6}. (Altschul et al., 1997). In the rare cases where BLASTP did not detect any homolog I used TBLASTN, which compared the protein query sequence against the nucleotide collection NCBI database translated in all reading frames. I used TBLASTN because the amino acid sequence of the protein encoded by one of the two overlapping frames (usually the one discovered more recently) may not be reported in the NCBI database.

Overall, I collected a total of 319 overlapping genes from 244 viral genomes (the number of genomes is lower than that of overlapping genes because some genomes contain more than one overlap). Each of them was classified as internal or terminal overlap. Internal overlap means that one gene is entirely within a second gene but in a different reading frame. Terminal overlap means that two genes overlap for part of their lengths, one upstream and one downstream. The combined overall length of overlapping genes was 163,050 nt.

For each viral genome, I extracted from the NCBI database the non-overlapping coding regions. They were composed of the sequences of non-overlapping genes, and, in cases where some genes partially overlap, of their non-overlapping regions. For viruses with segmented genomes, all segments were included in the calculations. Non-overlapping coding regions came from the 62 complete genomes that contain the 82 proven overlaps and from the 182 genomes that contain the respective 237 homologs (175 complete genomes and 7 partial genomes). The combined overall length of non-overlapping genes was 1,724,466 nt. All sequence data were collected in Supplementary File S1 (see Results).

2.2. Comparative analysis of overlapping and non-overlapping genes

I calculated the percent content of nucleotides, dinucleotides, amino acids, and synonymous codons in each overlapping gene and in the entire complement of non-overlapping genes in the viral genome. I also calculated the percent content of amino acids with high codon degeneracy (the 6-fold degenerate residues L, R, and S), medium codon degeneracy (the 4- and 3-fold degenerate residues A, G, P, T, V, and I), and low codon degeneracy (the 2- and 1-fold degenerate residues C, D, E, F, H, K, N, Q, Y, M, and W).

Overall, I compared a sample set of 319 overlapping genes with a control set of non-overlapping genes for 102 composition features (4 from nucleotides, 16 from dinucleotides, 20 from amino acids, 59 from synonymous codons, and 3 from amino acids grouped in accordance to codon degeneracy). I used the Wilcoxon signed-rank test for paired data (Siegel and Castellan, 1988) to detect the composition features showing a statistically significant difference between overlapping and non-overlapping genes (z-value > 4.41; two-tailed $P = 10^{-5}$; 319 degrees of freedom).

To overcome the problem that the dataset of non-overlapping genes was more than 10-fold larger than that of overlapping genes (1,724,466 vs. 163,050 nt), I randomly selected non-overlapping coding regions of the genomes that are of equal size to the overlapping set (510 nt from each genome for a total of 162,690 nt). By repeating this step 100 times, I obtained 100 resamplings of the non-overlapping set. I compared each resampling to the dataset of overlapping genes, assessing the statistical significance of the 102 composition features. I selected only the features showing a significant difference (z-value > 4.41) in more than 95 out of 100 resamplings.

2.3. Principal component analysis (PCA) of overlapping genes

The Wilcoxon test revealed that the nucleotide and amino acid composition of overlapping genes is significantly different from that of non-overlapping genes for 37 composition features (see Results). I used PCA (Hotelling, 1936; Morrison, 1976; Ringnér, 2008) to evaluate whether the observed differences were homogeneously distributed in individual overlapping genes, or if instead there were outliers with a highly atypical composition.

I calculated the difference between the percent content of the 37 composition features in each overlapping gene and the percent content of the same features in the respective complement of non-overlapping genes in the viral genome. I obtained a matrix of 319 rows (the number of overlaps) and 37 columns. The matrix was subjected to PCA, by using the OriginPro software (OriginLab, Northampton, MA). PCA summarized the information carried by 37 variables into 5 synthetic variables, that is the first (PC1), second (PC2), third (PC3), fourth (PC4), and fifth principal component (PC5). Using the first three PCs, I could represent the 319 overlapping genes of the dataset as a swarm of points in a three-dimensional map (see Results).

The search for overlapping genes with a highly atypical composition was carried out by applying to each of the five PCs the Rosner's test for detection of multiple outliers (Rosner, 1975) (https://contchart.com/outliers.aspx).

2.4. *Partial least squares-discriminant analysis (PLS-DA) and Fisher's linear discriminant analysis (LDA) of overlapping and non-overlapping genes*

After removing the 7 outliers detected by PCA (see Results), I compared a sample set of 312 overlapping genes with a control set of non-overlapping genes taken from 244 viral genomes. The first statistical method I used was the partial least squares-discriminant analysis (PLS-DA) (Brereton and Lloyd, 2014; Lee et al., 2018). The input data were collected into a matrix of 556 rows (312 overlapping and 244 non-overlapping genes) and 37 columns (the critical composition features I detected). The matrix was enlarged with a “dummy variable” having a value of -1 for overlapping genes and +1 for non-overlapping genes (38th column). This variable was the “outcome” variable on the Y-axis, while the 37 critical composition features were the “predictor” variables on the X-axes. PLS-DA first calculated *a*, the intercept on the Y-axis, and *b*, the regression coefficient for each X-axis. Using this linear regression function, I predicted Y and evaluated the accuracy of prediction. It was given by the number of overlapping genes with a predicted Y value below 0 added to the number of non-overlapping genes with a predicted Y value above 0. The sum was divided by the total number of observations (556) and multiplied by 100.

The accuracy of prediction was assessed with the validation test. The dataset was partitioned into: *i)* a training set composed of 75% of overlapping genes (234 out of 312) and of 75% of non-overlapping genes (183 out of 244); *ii)* a validation set composed of the remaining 25% of overlapping genes (78 out of 312) and of the remaining 25% of non-overlapping genes (61 out of 244). PLS-DA was carried out on the training set and the accuracy of prediction was tested on the validation set. By repeating this step 100 times, I obtained 100 random resamplings of the dataset. By subjecting each of them to PLS-DA, I could obtain the mean value of the accuracy of prediction. If the mean value and the accuracy of prediction from PLS-DA of the full dataset were both above 95%, I came to conclusion that this linear regression function had a discriminant power significantly high.

The other statistical method I used was LDA (Fisher, 1936; Lachenbruch and Goldstein, 1979). The input data were a matrix of 312 rows and 37 columns (the critical composition features I detected) and a matrix of 244 rows and 37 columns. The aim of LDA was to obtain a combination of a number of composition features yielding the best discrimination between overlapping and non-overlapping genes. Thus, I randomly selected a number of features ranging from 15 to 25 and I evaluated the accuracy of prediction (the percent of overlapping and non-overlapping genes correctly predicted as such) of the corresponding LDA.

The accuracy of prediction of each LDA was assessed with the validation test. By randomly resamplings the dataset, I obtained 100 training sets composed of 75% of overlapping and non-overlapping genes and 100 validation sets composed of the remaining 25% of overlapping and non-overlapping genes. LDA was carried out on each training set and the accuracy of prediction was tested on the respective validation set. If the mean value of the accuracy of prediction from validation sets and the accuracy of prediction from LDA of the full dataset were both above 95%, I came to conclusion that this linear function had a discriminant power significantly high.

Table 1
List of 17 composition features, sorted in accordance to a decreasing z-value, showing a significant enrichment in overlapping genes (z-value > 4.41; two-tailed P = 10^{-5}; 319 degrees of freedom) with respect to non-overlapping genes.

Composition feature	Mean % content in overlapping genes (sd)	Mean % content in non-overlapping genes (sd)	Mean % difference between overlap and non-overlap	z-value
Amino acids with high codon degeneracy	27.2 (4.0)	22.0 (2.7)	5.2	14.4
C	26.2 (5.8)	22.8 (5.2)	3.4	13.0
Arg	8.1 (3.2)	5.4 (1.1)	2.7	12.8
Ser	9.8 (2.8)	7.5 (1.8)	2.3	12.7
CG	4.7 (2.2)	3.4 (1.6)	1.3	12.5
CGA(Arg)	1.4 (1.1)	0.6 (0.3)	0.8	12.1
TCG(Ser)	1.4 (1.0)	0.6 (0.4)	0.8	11.9
TC	6.9 (2.2)	5.9 (1.7)	1.0	9.3
CC	7.2 (3.6)	5.7 (2.6)	1.5	9.2
AGC(Ser)	1.6 (1.0)	1.0 (0.5)	0.6	9.2
Gln	4.9 (1.9)	3.9 (0.9)	1.0	9.1
CCG(Pro)	1.3 (1.1)	0.7 (0.4)	0.6	8.9
CAG(Asn)	2.5 (1.3)	1.8 (0.7)	0.7	8.8
Pro	7.3 (3.8)	5.6 (1.6)	1.7	8.5
TCA(Ser)	2.3 (1.2)	1.7 (0.6)	0.6	7.7
AGG(Arg)	1.7 (1.1)	1.2 (0.6)	0.5	7.3
G	24.5 (4.3)	23.1 (3.1)	1.4	6.6

2.5. *Prediction of the genealogy of overlapping genes*

The genealogy of overlapping genes can be inferred by examining their phylogenetic distribution, under the assumption that the protein with the most restricted distribution is encoded by the *de novo* frame (Keese and Gibbs, 1992; Rancurel et al., 2009). The genealogy of 24 overlaps of the dataset is known from previous phylogenetic studies (see Table 1 in Sabath et al., 2012 and Table 1 in Pavesi et al., 2013). By extending it to the respective 87 homologs, I obtained a first set of 111 overlapping genes with a known ancestral and *de novo* frame.

Another approach to infer the genealogy of overlapping genes is the codon-usage method. It is based on the assumption that the ancestral frame, which has co-evolved with the other viral genes over a long period of time, has a distribution of synonymous codons significantly closer to that of the viral genome than the *de novo* frame (Pavesi et al., 2013). Analysis of the dataset with an improved version of the method (Pavesi, 2015) predicted the genealogy of 21 additional overlaps. By extending it to the respective 54 homologs, I obtained an additional set of 75 overlapping genes with a known ancestral and *de novo* frame.

The overlap polymerase/large envelope protein of *Hepatitis B virus* was examined separately. In accordance to the genealogy inferred by codon usage and indirect experimental evidence (Pavesi, 2015; Lauber et al., 2017), it was subdivided into two regions: *i)* a 5′ region in which the Pre-S domain of envelope (encoded by the ancestral frame) overlaps the spacer domain of polymerase (encoded by the *de novo* frame); *ii)* a 3' region in which the reverse transcriptase domain of polymerase (encoded by the ancestral frame) overlaps the S domain of envelope (encoded by the *de novo* frame). Extension of genealogy to 3 homologs yielded 4 other overlapping genes with a known ancestral and *de novo* frame.

Thanks to phylogenetic and codon-usage methods, I could predict the genealogy for more than half (46 out of 82) of the overlapping genes in the dataset (Supplementary Table S1).

2.6. Computer simulation of new overlapping genes encoding variants of the apoptin from Chicken anemia virus (CAV) and of the X protein from Borna disease virus (BDV)

The algorithm I developed first operated on the overlap capsid protein VP2/apoptin from CAV (Ac. number NC_001427). The apoptin (121 aa) is encoded by the *de novo* frame, which is shifted of one nucleotide 3' (+1) with respect to the ancestral frame. The algorithm randomly permuted the synonymous codons of the ancestral frame (e.g. permutation was between CGC(Arg) at codon position 3 and AGA(Arg) at codon position 7). Thus, permutation did not change the codon usage of the ancestral frame (e.g. CGA, CGC, and AGA for arginine remained the preferred codons over CGT, CGG, and AGG) as well as the amino acid sequence of the encoded capsid protein VP2.

However, permutation of synonyms did not affect 62 codon positions of the ancestral frame. The reason is that any synonymous change at these positions (e.g. CGA instead of CGC) would always cause an amino acid substitution in a few critical regions of the apoptin, as apoptin is encoded by the +1 frame. The critical regions (see Fig. 1 in Castro et al., 2018) are the following: *i)* a N-terminal apoptosis-inducing domain, formed by a proline-rich segment (20 aa; residues 9–28) and a (iso)leucine-rich segment (14 aa; residues 33–46); *ii)* a C-terminal apoptosis-inducing domain, formed by a bi-partite nuclear localization sequence (18 aa; residues 82–88 and 111–121) and a nuclear export sequence (9 aa; residues 97–105); *iii)* a threonine site at position 108, whose phosphorylation in transformed cells is crucial for nuclear accumulation of apoptin (Rohn et al., 2002).

Thus, permutation of synonyms in the ancestral frame was limited to 59 out of 121 codon positions. It generated a huge amount of new overlapping genes, half of which were discarded because of a +1 frame interrupted by termination codons. Selecting the remaining ones with the scoring rules given by PLS-DA and LDA, and using also a conservation criterion, yielded a subset of variants of the CAV apoptin (see Results).

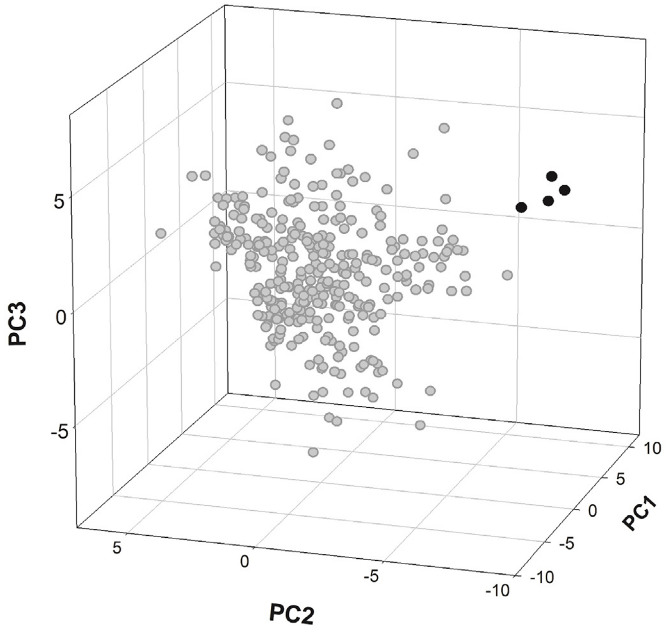


Fig. 1. Principal component analysis (PCA) of the 319 overlapping genes of the dataset. The three-dimensional map was obtained using the first (PC1), second (PC2), and third (PC3) principal component. They account for 26.7, 22.1, and 13.0% of the total amount of variation in the data, respectively. Black circles indicate the 4 homologs (from human, chimpanzee, woolly monkey, and long-fingered bat) of the overlap polymerase/X protein of *Hepatitis B virus*. They were classified as outlier from analysis of PC2 with the Rosner's test. Grey circles indicate the remaining 315 overlapping genes.

The algorithm I developed then operated on the overlap phosphoprotein/X protein from BDV (Ac. number NC_001607). The first 17 aa of the X protein are encoded by a non-overlapping region, while the remaining 70 aa are encoded by a *de novo* frame, which is shifted of two nucleotides 3' (+2) with respect to the ancestral frame. The algorithm randomly permuted the synonymous codons of the ancestral frame, preserving both its codon usage (e.g. CAG for glutamine remained the preferred codon over CAA) and the amino acid sequence of the encoded phosphoprotein. Permutation of synonyms affected all 70 codon positions of the ancestral frame, because little is known about the presence of functionally critical regions in the X protein. Permutation generated a huge amount of new overlapping genes, half of them were discarded because of a +2 frame interrupted by termination codons. Selection of the remaining ones with the scoring rules given by PLS-DA and LDA, and using also a conservation criterion, yielded a subset of variants of the X protein from BDV (see Results).

3. Results

3.1. Creation of a dataset of 319 homologous overlapping genes and preliminary analyses

The search for homologs in the NCBI database led to detection of only one homolog for 30 out of 82 overlaps in the dataset. In the remaining cases, it led to detection of a mean number of 2.9 homologs per overlap, with a standard deviation (sd) of 2.6. In detail, I found a large number of homologs in the following cases: movement protein/replicase of *Turnip yellow mosaic virus* (19 homologs); p22 protein/p19 protein of *Tomato bushy stunt virus* (9 homologs); polyprotein (core protein)/F protein of *Hepatitis C virus* (8 homologs); RdRp (2a)/2b protein of *Spinach latent virus* (7 homologs).

The dataset contains a total of 319 overlapping genes with a mean length of 511 nt (sd = 476 nt). The high standard deviation is due to a wide length distribution, ranging from 126 nt (the homolog of the overlap bel protein/bet protein of *Puma feline foamy virus*) and 2844 nt (p130 protein/replicase of *Providence virus*). The combined overall length of overlapping genes was 163,050 nt, while that of non-overlapping genes was more than 10-fold higher (1,724,466 nt).

I collected all nucleotide and amino acid sequence data in Supplementary File S1. Section A of the file contains a list of the 82 overlaps with the respective 237 homologs. Section B reports a list of the 244 virus species (or isolates of a given virus species) that contain the 319 overlaps of the dataset. For each virus, it specifies the number of overlaps, their length, the overall length of non-overlapping genes, and the length of the genome. Section B also reports the classification of each virus in accordance to the genome type. The 319 overlaps are distributed as follows: 162 overlaps from 125 ssRNA + viruses, 65 overlaps from 47 ssRNA-viruses, 19 overlaps from 12 ssRNA-RT viruses, 13 overlaps from 13 dsRNA viruses, 47 overlaps from 38 ssDNA viruses, 5 overlaps from 5 dsDNA viruses, and 8 overlaps from 4 dsDNA-RT viruses.

For each overlapping gene, section C contains the following information: I) accession number in the NCBI database; II) name of virus species, family, and genus; III) name of the overlapping gene; IV) nucleotide sequence of the upstream overlapping frame; V) amino acid sequence of the protein encoded by the upstream overlapping frame; VI) nucleotide sequence of the downstream overlapping frame, shifted of one or two nucleotides 3' with respect to the upstream frame; VII) amino acid sequence of the protein encoded by the downstream overlapping frame; VIII) nucleotide sequence of the entire complement of non-overlapping genes in the viral genome; IX) amino acid sequence of the proteins encoded by the entire complement of non-overlapping genes.

Using the data reported in section B, I calculated the density of overlapping genes in each virus species. It was the ratio between the length of overlapping genes and the length of the genome, multiplied by 100. I found that about two-thirds of viruses (150 out of 244) have a density of overlapping genes less than 10% (from 1.1% in ebolaviruses to 9.6% in panicoviruses). The highest density (50%) was found in *Hepatitis B virus*, followed by *Providence virus* (44%), *Maize chlorotic mottle virus* (34%), *Hibiscus chlorotic ringspot virus* (33%), and tymoviruses (from 21.8% in *Erysimum latent virus* to 35% in *Kennedya yellow mosaic virus*).

I then examined the 244 virus species of the dataset to determine whether there was a relationship between the length of their genomes and of their overlapping genes. Using the Spearman rank correlation coefficient, I found a weak, albeit significant, negative correlation (rho = -0.31; t-Student = 5.36; P = 10^{-4}).

Finally, I found a statistically significant prevalence of the internal overlaps (185 out of 319 overlaps) on the terminal ones (134 out of 319 overlaps) (chi-square = 15.67; P = 10^{-5}). It mainly depends on ssRNA-viruses, in which the number of internal overlaps (53 out of 65) was significantly higher than that of terminal overlaps (12 out of 65) (chi-square = 17.38; P = 10^{-5}). In the other 6 classes of viruses, the difference was statistically not significant (chi-square < 3.84).

3.2. Overlapping genes differ significantly from non-overlapping genes for 37 composition features

The Wilcoxon test revealed that the nucleotide and amino acid composition of overlapping genes is significantly different (z-value > 4.41; two-tailed P = 10^{-5}; 319 degrees of freedom) from that of non-overlapping genes for one third of the examined composition features (37 out of 102). Indeed, all these features showed a z-value higher than the cutoff of significance (4.41) in more than 95 out of 100 resamplings of the non-overlapping set (see Methods).

Table 1 shows a list of 17 features, sorted in accordance to a decreasing z-value, in which overlapping genes exhibited a significant enrichment with respect to non-overlapping genes. For instance, overlapping genes have a mean percent content in amino acids with high codon degeneracy (Arg, Leu, and Ser) of 27.2%, which is 5.2% greater than that in non-overlapping genes (z-value = 14.4). Considering the other z-values > 9.0, overlapping genes are also enriched in the nucleotide C, the dinucleotides CG, TC, and CC, the amino acids Arg, Ser, and Gln, and the synonymous codons CGA(Arg), TCG(Ser) and AGC(Ser). These enrichments are clearly linked. Regarding amino acids that are enriched, Arg is encoded by codons rich in CG (among which CGA, which is also enriched); likewise, Ser is largely encoded by TCG, which is a combination of the enriched dinucleotides TC and CG.

Table 2 shows a list of 20 features, sorted in accordance to a decreasing z-value, in which overlapping genes exhibited a significant depletion with respect to non-overlapping genes. Considering the z-values > 10.0, overlapping genes are depleted in the nucleotide T, the dinucleotides TA, AT, TT, and TG, the amino acids Tyr, Val, Phe, and Ile, the synonymous codons TAT(Tyr), TTT(Phe), ATT(Ile), and GTT(Val), and the amino acids with low codon degeneracy. These depletions are correlated each other. Regarding amino acids that are depleted, Tyr, Phe, and Ile are encoded by codons rich in A and T (among which TAT, TTT, and ATT, which are also depleted). Depletion in TA and TG has a clear biological meaning, as it reduces the probability of occurrence of termination codons (TGA, TAG, and TAA) in overlapping genes.

3.3. Principal component analysis (PCA) of overlapping genes revealed the presence of 7 outliers

PCA summarized the information carried by 37 variables (the critical composition features I detected) into 5 synthetic variables, that is the first (PC1), second (PC2), third (PC3), fourth (PC4), and fifth principal component (PC5). They accounted, respectively, for 26.7, 22.1, 13.0, 7.0, and 6.3% of the total variation in the data. Taken together, the first 5 PCs summarized 75% of the total variation, i.e. reduction from 37 to 5 variables resulted in a loss of information of 25%. Using the first 3 PCs, the 319 overlapping genes of the dataset were represented as a swarm of points into a three-dimensional map (Fig. 1).

Table 2
List of 20 composition features, sorted in accordance to a decreasing z-value, showing a significant depletion in overlapping genes (z-value > 4.41; two-tailed P = 10^{-5}; 319 degrees of freedom) with respect to non-overlapping genes.

Composition feature	Mean % content in overlapping genes (sd)	Mean % content in non-overlapping genes (sd)	Mean % difference between overlap and non-overlap	z-value
Tyr	2.2 (1.3)	3.6 (0.7)	-1.4	12.8
TA	3.6 (1.5)	5.3 (1.8)	-1.7	12.6
AT	5.7 (1.9)	7.1 (1.8)	-1.4	12.2
Amino acids with low codon degeneracy	36.9 (6.1)	41.2 (3.9)	-4.3	12.1
TAT(Tyr)	1.1 (0.9)	2.0 (0.8)	-0.9	11.7
T	21.6 (4.3)	25.3 (3.1)	-3.7	11.5
TTT(Phe)	1.3 (1.1)	2.3 (0.9)	-1.0	11.2
Val	5.1 (2.1)	6.5 (1.3)	-1.4	10.8
Phe	3.0 (1.7)	4.1 (0.8)	-1.1	10.6
TT	5.2 (2.3)	6.9 (1.8)	-1.7	10.4
TG	6.0 (1.8)	7.1 (1.1)	-1.1	10.4
ATT(Ile)	1.4 (1.1)	2.2 (0.8)	-0.8	10.4
Ile	4.3 (2.0)	5.6 (1.3)	-1.3	10.3
GTT(Val)	1.3 (1.0)	2.0 (0.8)	-0.7	10.2
AAT(Asn)	1.9 (1.3)	2.7 (1.1)	-0.8	10.0
GAT(Asp)	2.2 (1.5)	2.9 (1.1)	-0.7	8.7
Asp	4.2 (1.8)	5.1 (0.8)	-0.9	8.5
Asn	3.9 (1.6)	4.7 (1.2)	-0.8	8.3
GT	4.5 (1.5)	5.1 (1.0)	-0.6	7.8
AAA(Lys)	2.4 (1.7)	3.2 (1.4)	-0.8	7.6

Application of the Rosner's test for detection of multiple outliers to the first 3 PCs did not reveal any outliers in PC1 and PC3, but 4 outliers in PC2 (P = 0.02). They were the 4 homologs of the overlapping gene of *Hepatitis B virus* that encodes the RNase domain of polymerase and the X protein (see black circles in Fig. 1). The atypical composition of this overlap mainly depends on: *i)* an extremely higher content in the nucleotide C (37.3%) and the dinucleotide CG (9.6%), with respect to the non-overlapping genome counterpart (23.1% and 2.1%, respectively); *ii)* an extremely lower content in amino acids with low codon degeneracy (24.0%), with respect to the non-overlapping genome counterpart (40.8%).

Application of the Rosner's test to the remaining PCs led to detection of 3 outliers in PC4 (P = 0.01). They were the 3 homologs of the overlapping gene that encodes the capsid protein VP2 and the nucleocapsid protein VP1 in gyroviruses. Their atypical composition mainly depends on an extremely higher content in arginine (22.5%) and the codon CGA(Arg) (6.4%), with respect to the non-overlapping genome counterpart (5.2% and 0.8%, respectively).

3.4. Asymmetric evolution is the prevailing pattern in overlapping genes

Asymmetric evolution means that the protein encoded by one frame (usually the *de novo* frame) is significantly more variable than that encoded by the other frame. A previous analysis of 65 pairs of homologous overlaps revealed that half of them undergo asymmetric evolution (Pavesi, 2019).

The increased content in homologs of the present dataset allowed a more exhaustive analysis. Compared to the previous study, I could assess the pattern of symmetric/asymmetric evolution by examining a number of homologs three-fold greater (208 vs. 65). Using Clustal Omega (Sievers and Higgins, 2014), I first aligned the two proteins

Table 3
List of the 5 overlapping genes in which analysis of a higher number of homologs revealed that they undergo asymmetric evolution (chi-square value $>$ 3.84; 1 degree of freedom)..

Genome ac. number[a]	Virus species[b]	Overlapping gene (protein products)	Number of conserved and non-conserved amino acid positions in the alignment[c]	Chi-square value	Most variable protein
NC_003809 (EU919669, NC_003568, NC_003834, NC_003547, NC_009538, NC_011809)	Spinach latent virus (Lilac ring mottle virus, Elm mottle virus, Tulare apple mosaic virus, Tomato necrotic streak virus, Citrus leaf rugose virus, Citrus variegation virus, Asparagus virus 2)	RNA-dependent RNA polymerase (2a)/2b protein	2a: 7, 36 (16.3%) 2b: 19, 24 (44.2%)	6.67 (P = 0.01)	RNA-dependent RNA polymerase (2a)
NC_004146 (NC_004142, KU754529)	Flock house virus (Boolarra virus, Newington virus)	RNA-dependent RNA polymerase(A)/B2 protein	A: 53, 46 (53.5%) B2: 72, 28 (72%)	6.49 (P = 0.01)	RNA-dependent RNA polymerase (A)
NC_001747 (NC_006265, NC_034265, NC_016038, NC_031747, NC_008249, NC_030225)	Potato leafroll virus (Carrot red leaf virus, Tobacco virus 2, Brassica yellows virus, White clover mottle virus, Chickpea chlorotic stunt virus, Pepo aphid-borne yellows virus)	capsid protein (P3)/movement protein (P4)	P3: 38,52 (42.2%) P4: 19, 71 (21.1%)	8.32 (P = 0.004)	movement protein (P4)
NC_004674 (NC_000869, KM978186, NC_003708, NC_030310)	East African cassava mosaic virus (Tomato yellow leaf curl Thailand virus, Euphorbia yellow leaf curl virus, Watermelon chlorotic stunt virus, Cotton yellow mosaic virus)	transcriptional activator (TrAP, AC2 protein)/replication enhancer (Ren, AC3 protein)	AC2: 28, 58 (32.6%) AC3: 46, 51 (52.9%)	6.49 (P = 0.01)	transcriptional activator (TrAP, AC2 protein)
NC_003977 (HPBVCG, NC_028129, NC_020881)	Hepatitis B virus isolate USA (Chimpanzee hepatitis B virus, Woolly monkey hepatitis B virus, Long-fingered bat hepatitis B virus)	X protein/polymerase (P)	X: 36, 40 (47.4%) P: 50, 27 (64.9%)	4.11 (P = 0.04)	X protein

[a] Parenthesis indicates the ac. number of the homologs.
[b] Parenthesis indicates the source of the homologs.
[c] Parenthesis indicates the percent of the amino acid identity.

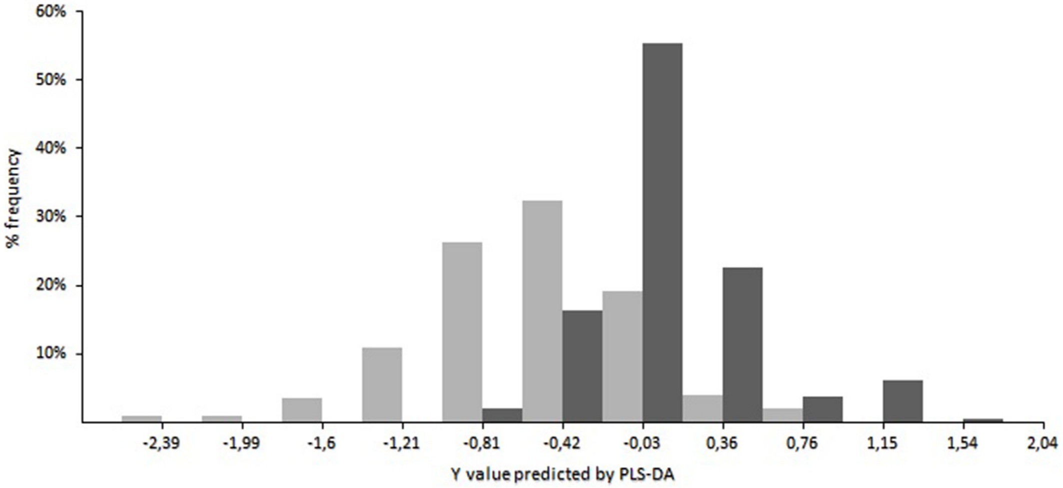


Fig. 2. Histogram of the distribution of the Y value predicted by PLS-DA in 312 overlapping genes (grey columns) and in 244 non-overlapping genes (black columns). The linear regression function consists of the intercept on the Y-axis *a* (value = 1.617) and of 23 values of *b*, the regression coefficient for each composition feature. The values of the regression coefficient are as follows: 0.001 for G, -0.035 for AT, -0.046 for TT, -0.036 for TG, 0.028 for TC, -0.116 for GT, 0.002 for CG, 0.068 for Ser, 0.113 for Val, -0.126 for Gln, 0.155 for Asp, 0.200 for Tyr, 0.088 for Phe, -0.080 for amino acids with high codon degeneracy, -0.164 for CGA(Arg), -0.176 for TCA(Ser), -0.221 for TCG(Ser), -0.158 for AGC(Ser), -0.086 for CCG(Pro), 0.010 for GTT(Val), -0.013 for AAT(Asn), -0.021 for GAT(Asp), and 0.113 for ATT(Ile). A high percentage of overlapping genes (94.9%) were correctly classified as overlap, as their predicted Y value was below 0 (from -2.778 to -0.005). A high percentage (98.4) of non-overlapping genes were correctly classified as non-overlap, as their predicted Y value was above 0 (from 0.013 to 1.936).

encoded by each overlap with the respective homologs. Whenever present, I then manually removed the protein regions with large gaps.

Consider for example the overlap capsid protein P3/movement protein P4 of *Potato leafroll virus*, which encodes two proteins having a length of 156 aa. Manual inspection of the alignments of P3 and P4 with the respective 6 homologs led to removal of the largely gapped N- and C-terminal regions. Thus, the occurrence of symmetric or asymmetric evolution was tested on the central un-gapped region of 90 aa. Using the chi-square test (cutoff > 3.84; $P < 0.05$; one degree of freedom), the numbers of conserved and non-conserved amino acid positions in the alignment of P3 (38 and 52, respectively) were compared to those in the alignment of P4 (19 and 71, respectively). A chi-square value of 8.32 ($P = 0.004$) indicated that this overlap undergoes asymmetric evolution, as the number of amino acid substitutions in P4 is significantly higher than that in P3.

Chi-square test confirmed the pattern of asymmetric evolution found previously in 32 overlaps (see Table 1 in Pavesi, 2019). Unlike the previous study, it revealed that 5 more overlapping genes evolve in accordance to the asymmetric model (Table 3). Thus, the number of overlaps evolving asymmetrically increased from 32 to 37 (57% of the total), while that of overlaps evolving symmetrically decreased from 33 to 28 (43% of the total). Although the percent difference (57% vs. 43%) was statistically not significant (chi-square value = 1.97; $P = 0.16$), asymmetric evolution is now the prevailing pattern in viral overlapping genes.

3.5. Partial least squares-discriminant analysis (PLS-DA) and Fisher's linear discriminant analysis (LDA) separated overlapping genes from non-overlapping genes with an accuracy higher than 95%

Having found that PCA has a poor ability to separate overlapping genes from non-overlapping genes (see, respectively, the grey and black circles in Supplementary Figs. S1) and I explored the dataset using PLS-DA and LDA, that is supervised multivariate statistical methods that maximize the variance between groups and minimize the variance within groups.

Analysis with PLS-DA of a matrix of 556 rows (312 overlapping genes after exclusion of 7 outliers and 244 non-overlapping genes) and 38 columns (the 37 critical composition features I detected and a dummy variable assigning -1 to overlapping genes and +1 to non-overlapping genes) yielded a linear regression function consisting of *a*, the intercept on the Y axis, and 37 values of *b*, the regression coefficient.

This function assigned a predicted Y value below 0 to 296 out of 312 overlapping genes (error of classification = 5.1%) and a predicted Y value above 0 to 241 out 244 non-overlapping genes (error of classification = 1.2%). The total error was 3.4% (19 misclassifications out of 556) with an accuracy of prediction of 96.6%. The accuracy of prediction of the linear regression function was assessed with the validation test (see Methods). When applied to 100 validation sets, it yielded for overlapping genes a mean error of classification of 6.3% and for non-overlapping genes of 3.1%. The mean total error of classification was 4.9% with a mean accuracy of prediction of 95.1%.

To obtain a linear regression function with a highest discriminant power, especially for overlapping genes, I randomly selected a number of critical composition features ranging from 15 to 25 and I evaluated the accuracy of prediction of the corresponding PLS-DA. The best performance was given by a PLS-DA accounting for 23 composition features (1 from nucleotides, 6 from dinucleotides, 7 from amino acids, and 9 from synonymous codons). When applied to 100 validation sets, it yielded a mean error of classification of 5.4% for overlapping genes and of 3.2% for non-overlapping genes. The mean total error of classification was 4.5% with a mean accuracy of prediction of 95.5%. When applied to the full dataset, this linear regression function misclassified 5.1% of overlapping genes and 1.6% of non-overlapping genes. The total error was 3.6% with an accuracy of prediction of 96.4%. The strong discriminant power of this linear regression function is evident from the distribution of the predicted Y value in overlapping (grey columns) and non-overlapping genes (black columns) (Fig. 2).

The other statistical method I used was LDA. The input data were a matrix of 312 rows (the number of overlaps after exclusion of 7 outliers) and 37 columns (the critical composition features I detected) and a matrix of 244 rows (the number of the genome complements of non-overlapping genes) and 37 columns. I randomly selected a number of features ranging from 15 to 25 and I evaluated the accuracy of prediction of the corresponding LDA. The performance of each LDA was then assessed with the validation test (see Methods).

The best performance was given by a LDA accounting for 21 composition features (2 from nucleotides, 4 from dinucleotides, 8 from amino acids, and 7 from synonymous codons). When applied to 100

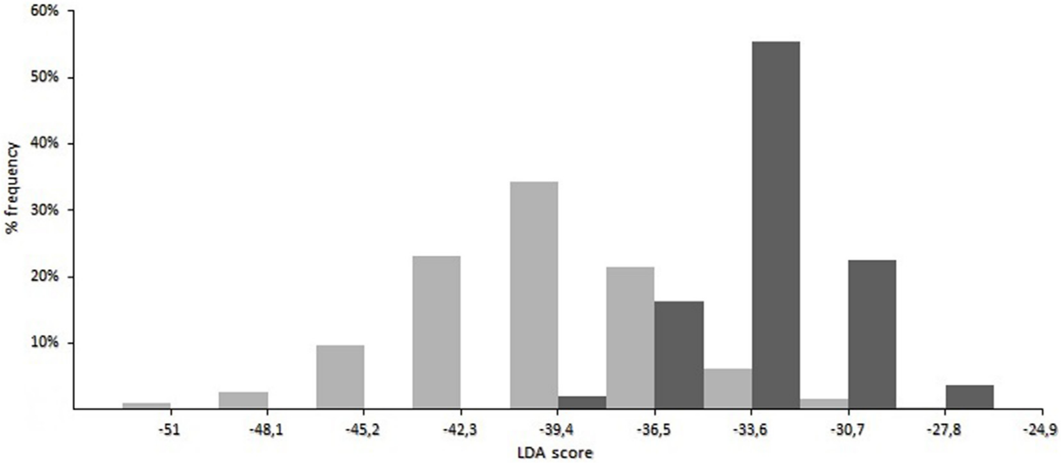


Fig. 3. Histogram of the distribution of the LDA score in 312 overlapping genes (grey columns) and in 244 non-overlapping genes (black columns). The linear function, accounting for 21 composition features, consists of 21 coefficients. Their values are as follows: 0.361 for T, -0.255 for G, -0.829 for AT, -0.543 for TA, -0.701 for TG, -0.105 for CG, 0.217 for Arg, -0.171 for Pro, -0.861 for Asn, -1.084 for Gln, 1.320 for Asp, 1.428 for Tyr, -0.884 for amino acids with high codon degeneracy, -0.164 for amino acids with low codon degeneracy, -1.228 for CGA(Arg), 0.018 for AGG(Arg), -1.014 for TCG(Ser), -0.656 for CCG(Pro), 0.328 for AAA(Lys), 0.704 for AAT(Asn), and 0.115 for ATT(Ile). With a discriminant score of -35.31, a high percentage (96.5%) of overlapping genes were correctly classified as overlap (their score ranged from -53.91 to -35.36). With the same score, a high percentage (97.1%) of non-overlapping genes were correctly classified as non-overlap (their score ranged from -35.31 to 0.84).

validation sets, it yielded a mean error of classification of 3.3% for overlapping genes and of 4.6% for non-overlapping genes. The mean total error of classification was 3.9% with a mean accuracy of prediction of 96.1%. When applied to the full dataset, this linear function misclassified 3.5% of overlapping genes and 2.9% of non-overlapping genes. The total error was 3.2% with an accuracy of prediction of 96.8%. The strong discriminant power of this linear function can be appreciated by examining the distribution of the LDA score in overlapping (grey columns) and non-overlapping genes (black columns) (Fig. 3).

I summarized the performance of PLS-DA and LDA into a two-dimensional map in which the predicted Y value in overlapping and non-overlapping genes was plotted against the respective LDA score (Fig. 4). The grey circles into part A of the map are the 294 overlaps correctly classified by both methods (94.2% of the total). The black circles into part C are the 237 non-overlaps correctly classified by both methods (97.1% of the total).

Finally, I found that the performance of both PLS-DA and LDA is not affected by the fact that the dataset includes representatives from all Baltimore classes of viruses (e.g. ssRNA+, ssRNA-, ssDNA, and dsDNA viruses). When the validation test was applied to the 259 overlaps from RNA viruses, the linear regression function accounting for 23 composition features showed a total mean error of classification of 3.2%. When the same test was applied to the 162 overlaps from ssRNA + viruses, it yielded a mean total error of classification of 2.8%. Similarly, the LDA function accounting for 21 composition features showed a mean total error of classification of 4.8%, when the validation test was applied to the 259 overlaps from RNA viruses, and of 2.4% when the same test was applied to the 162 overlaps from ssRNA + viruses.

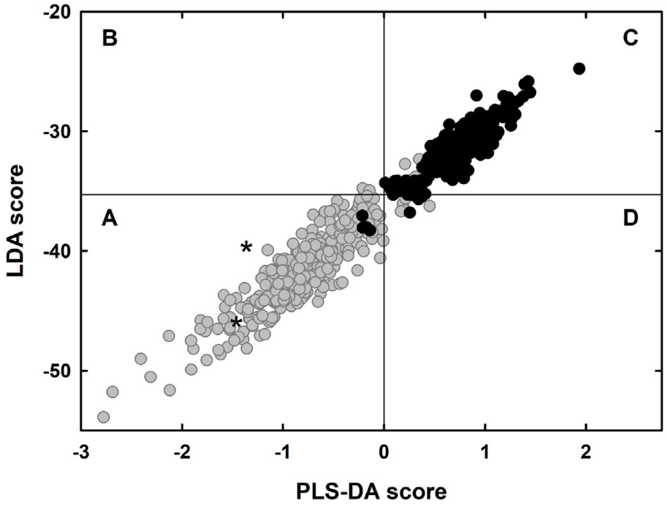


Fig. 4. Map of overlapping genes (grey circles) and non-overlapping genes (black circles), in which the Y value predicted by PLS-DA is plotted against the respective LDA score. Part A of the map includes 294 overlaps correctly classified by PLS-DA and LDA, and 4 non-overlaps misclassified by both methods. Part B includes 2 overlaps correctly classified by PLS-DA, but misclassified by LDA. Part C includes 237 non-overlaps correctly classified by PLS-DA and LDA, and 9 overlaps misclassified by both methods. Part D includes 7 overlaps correctly classified by LDA, but misclassified by PLS-DA, and 3 non-overlaps correctly classified by PLS-DA, but misclassified by LDA. Asterisks indicate the two new potential overlapping genes detected in the genome of the isolate Wuhan-Hu-1 of SARS-CoV-2: the overlaps 3a protein/hypothetical protein and the overlap nucleocapsid protein/hypothetical protein.

3.6. Application of the model to a sample of 8 putative overlapping genes

Using the intercept and the 23 regression coefficients given by PLS-DA (see legend of Fig. 2) and the 21 coefficients given by LDA (see legend of Fig. 3), I applied the model to a sample set of 8 overlapping genes that were previously classified as putative, due to poor experimental evidence about their expression (Supplementary Table S2 from Pavesi et al., 2018). Five of them were classified as "*bona fide*" overlapping genes, because they were predicted as such by both PLS-DA and LDA. They were the overlaps p17/capsid protein p71 of *Dendrolimus punctatus tetravirus* (NC_005899), GP2'/GP3′ of *Simian hemorrhagic fever virus* (NC_003092), ORF2/ORF3 of *Mushroom bacilliform virus* (NC_001633), NS2A-NS2B/NS2AN-FIFO of *Culex flavivirus* (NC_008604), and RdRp/36.9 kDa protein of *Rice ragged stunt virus* (NC_003771). Two overlaps (capsid protein/23 kDa NAB protein of *Indian citrus ringspot virus*, NC_003093 and ORF1/ORF2 of *Mushroom bacilliform virus*, NC_001633) remain putative, because they were positive to only one of the two methods. Being negative to both methods, the overlap GP2'/E' of *Simian hemorrhagic fever virus* (NC_003092) was classified as unreliable.

3.7. Application of the model to the viral genome of SARS-CoV-2

The present dataset contains two overlapping genes (3a protein/3b protein and nucleocapsid protein/9b protein) from human SARS-Cov (Ac. number NC_004718) and the 2 respective homologs from bat SARS-Cov (Ac. number DQ412042) (see overlaps from #51 to #54 in Supplementary File S1). Thus, it was important to examine the genome sequences of the new coronavirus SARS-CoV-2, the etiological agent of the current pneumonia outbreak in the world (Zhou et al., 2020). I analyzed the genome sequence of the isolate Wuhan-Hu-1 of SARS-CoV-2 (NC_045512), which shows a nucleotide identity of 80% with human and bat SARS-Cov and of 96% with the closely related bat coronavirus RaTG13 (Ac. number MN996532).

Application of the model to SARS-CoV-2 pointed out a strict conservation of the overlap nucleocapsid protein/9b protein. Indeed, the 9b protein of SARS-CoV-2 (unannotated in NC_045512, but encoded from nt 28,284 to 28,574) showed an identity of 71% with the homologs from human and bat SARS-Cov and of 93% with the homolog from bat coronavirus RaTG13 (unannotated in MN996532, but encoded from nt 28,250 to 28,543).

Interestingly, the model pointed out that the overlap 3a protein/3b protein of SARS-CoV-2 is conserved in bat coronavirus RATG13, but it is deeply different from that of human and bat SARS-CoV. Indeed, the 3b protein of SARS-CoV-2 (not annotated in NC_045512, but encoded from nt 25,457 to 25,579) is three-fold shorter (41 aa) than that of human and bat SARS-CoV (154 and 114 aa, respectively). In addition, it shows with them an identity of only 22.5 and 28%, respectively.

Application of the model led to identification in SARS-CoV-2 of two new potential overlapping genes. They were predicted as such by both PLS-DA and LDA (see asterisks in Fig. 4). In the first case, a +1 overlapping frame from nt 28,734 to 28955 is entirely within the region that encodes residues 155–229 of nucleocapsid protein. It encodes a hypothetical protein having a length of 73 aa (Fig. 5A). Using TBLASTN, I found that this hypothetical protein shows an identity of 90% with the homolog from bat coronavirus RaTG13 (nt 28700–28918), and of 77% and 73% with the homologs (3 aa shorter) of human and bat SARS-Cov (nt 25,583–28,792 and 28,544–28,753, respectively).

In the other case, a +2 overlapping frame from nt 25524 to 25697 is entirely within the region that encodes residues 44 to 102 of 3a protein. This frame encodes a hypothetical protein having a length of 57 aa (Fig. 5B). Using TBLASTN, I found that this overlapping frame is a peculiar feature of SARS-CoV-2. Indeed, both the homologous genome region of bat coronavirus RaTG13 (nt 25,494–25,667) and that of human and SARS-CoV (nt 25,399–25,572 and nt 25,338–25,511, respectively) appear to be interrupted by several termination codons.

3.8. Fisher's linear discriminant analysis (LDA) separated ancestral from the de novo frames with an accuracy close to 100%

Thanks to the phylogenetic and codon-usage methods, I could predict the genealogy for more than half (46 out of 82) of the overlapping genes in the dataset (Supplementary Table S1). I subdivided the overlap polymerase/large envelope protein of *Hepatitis B virus* into two regions, each with its own genealogy (see Methods). I thus obtained a dataset of 47 overlapping genes, enriched by the respective 147 homologs, with a known ancestral and *de novo* frame.

I subdivided the dataset into two groups. In the first (36 overlaps and the respective 90 homologs) the *de novo* frame is shifted of one nucleotide 3' (+1) with respect to the ancestral frame (Supplementary File S2). In the other (11 overlaps and the respective 58 homologs) the *de novo* frame is shifted of two nucleotides 3' (+2) with respect to the ancestral frame (Supplementary File S3).

Supplementary Files S2 and S3 report for each overlap the following information: I) the accession number in the NCBI database; II) the name of virus species, family, and genus; III) the name of the overlapping gene; IV) the nucleotide sequence of the ancestral frame; V) the amino acid sequence of the protein encoded by the ancestral frame; VI) the nucleotide sequence of the *de novo* frame; VII) the amino acid sequence of the protein encoded by the *de novo* frame.

I analyzed the two groups of overlaps separately, with the aim to detect critical composition differences between ancestral and +1 *de novo* frames in one case, and between ancestral and +2 *de novo* frames in the other. I calculated the percent content in amino acids, synonymous codons, and amino acids grouped in accordance to codon

```
(A)
aat gct gca atc gtg cta caa ctt cct caa gga aca aca ttg cca aaa ggc ttc tac
 N   A   A   I   V   L   Q   L   P   Q   G   T   T   L   P   K   G   F   Y
  M   L   Q   S   C   Y   N   F   L   K   E   Q   H   C   Q   K   A   S   T

gca gaa ggg agc aga ggc ggc agt caa gcc tct tct cgt tcc tca tca cgt agt cgc
 A   E   G   S   R   G   G   S   Q   A   S   S   R   S   S   S   R   S   R
  Q   K   G   A   E   A   A   V   K   P   L   L   V   P   H   H   V   V   A

aac agt tca aga aat tca act cca ggc agc agt agg gga act tct cct gct aga atg
 N   S   S   R   N   S   T   P   G   S   S   R   G   T   S   P   A   R   M
  T   V   Q   E   I   Q   L   Q   A   A   V   G   E   L   L   L   L   E   W

gct ggc aat ggc ggt gat gct gct ctt gct ttg ctg ctg ctt gac aga ttg aac
 A   G   N   G   G   D   A   A   L   A   L   L   L   L   D   R   L   N
  L   A   M   A   V   M   L   L   L   L   C   C   C   L   T   D   *

(B)
gga tgg ctt att gtt ggc gtt gca ctt ctt gct gtt ttt cag agc gct tcc aaa atc
 G   W   L   I   V   G   V   A   L   L   A   V   F   Q   S   A   S   K   I
     M   A   Y   C   W   R   C   T   S   C   C   F   S   E   R   F   Q   N

ata acc ctc aaa aag aga tgg caa cta gca ctc tcc aag ggt gtt cac ttt gtt tgc
 I   T   L   K   K   R   W   Q   L   A   L   S   K   G   V   H   F   V   C
H   N   P   Q   K   E   M   A   T   S   T   L   Q   G   C   S   L   C   L

aac ttg ctg ttg ttg ttt gta aca gtt tac tca cac ctt ttg ctc gtt gct gct ggc
 N   L   L   L   L   F   V   I   V   Y   S   H   L   L   L   V   A   A   G
Q   L   A   V   V   V   C   N   S   L   L   T   P   F   A   R   C   C   W

ctt gaa
 L   E
P   *
```

Fig. 5. Nucleotide and amino acid sequence of the 2 new potential overlapping genes detected in the viral genome of SARS-CoV-2 (Ac. number NC_045512). (A) Overlap nucleocapsid protein/hypothetical protein: the nucleotide sequence (from nt 28733 to 28597) encodes the region of nucleocapsid protein spanning residues 155–229. The +1 overlapping frame (from nt 28734 to 28955) encodes a hypothetical protein (underlined characters) with a length of 73 amino acids. (B) Overlap 3a protein/hypothetical protein: the nucleotide sequence (from nt 25522 to 25698) encodes the region of 3a protein spanning residues 44–102. The +2 overlapping frame (from nt 25524 to 25697) encodes a hypothetical protein (underlined characters) with a length of 57 amino acids.

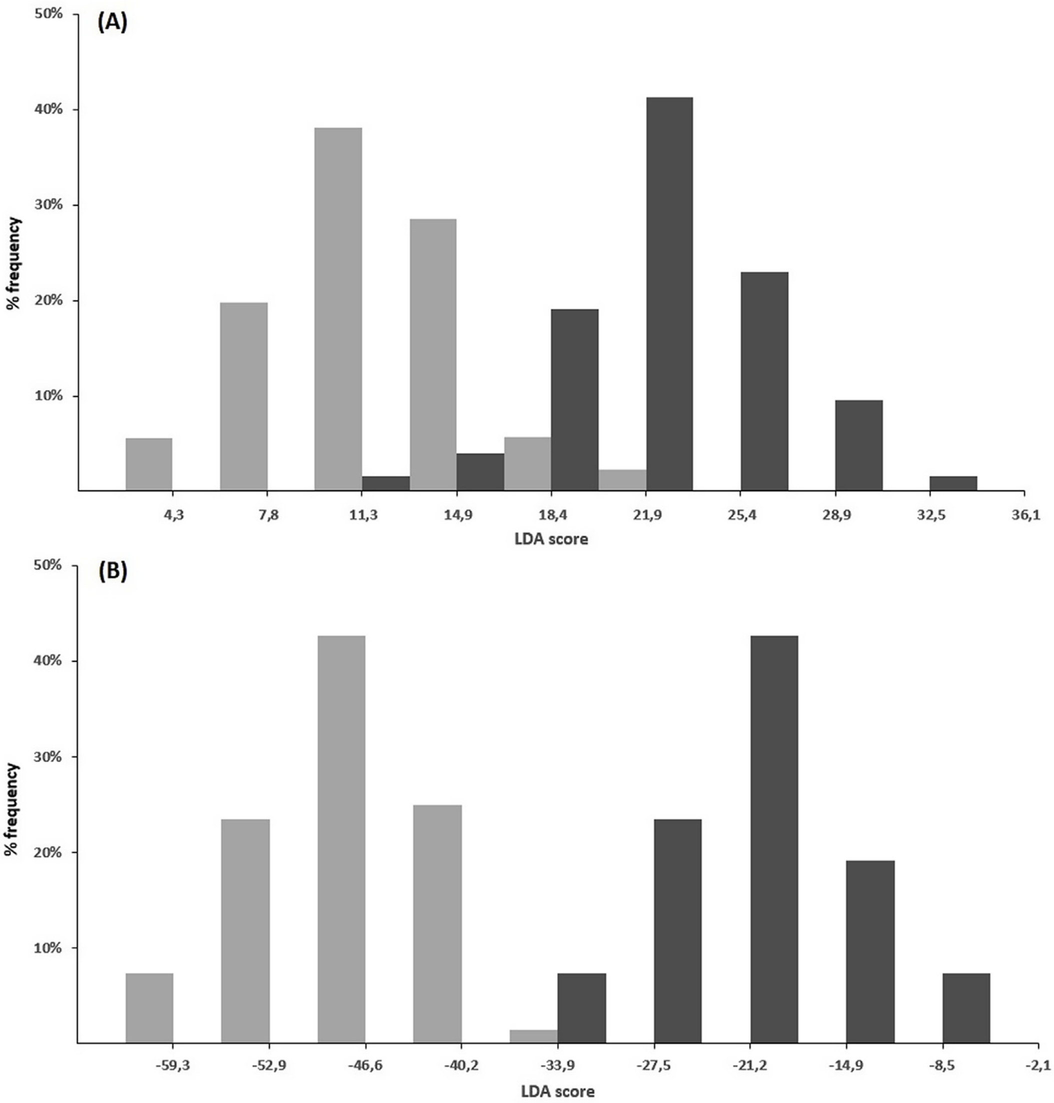


Fig. 6. (A) Histogram of the distribution of the LDA score in 126 ancestral frames (black columns) and in the respective +1 *de novo* frames (grey columns). With a discriminant score of 17.20, a high percentage (96.8%) of ancestral frames were correctly classified as ancestral (their score ranged from 17.24 to 35.98). With the same score, a high percentage (97.6%) of the +1 *de novo* frames were correctly classified as *de novo* (their score ranged from 0.77 to 16.78). (B) Histogram of the distribution of the LDA score in 68 ancestral frames (black columns) and in the respective +2 *de novo* frames (grey columns). With a discriminant score of -34.98, all the ancestral frames and all the +2 *de novo* frames were correctly classified as ancestral and *de novo*, respectively.

degeneracy in each ancestral frame and in the respective *de novo* frame. I used the Wilcoxon test to detect the composition features showing a statistically significant difference (z-value > 4.41; two-tailed $P = 10^{-5}$) between ancestral and *de novo* frames.

When applied to the first group (126 overlaps), the Wilcoxon test revealed that the +1 *de novo* frames differ significantly from the ancestral ones for 25 composition features. Section A of Supplementary Table S2 shows a list of 14 features in which the +1 *de novo* frames exhibited a significant enrichment with respect to the ancestral ones. Section B shows a list of 11 features in which the +1 *de novo* frames exhibited a significant depletion.

Using these composition features as input data for LDA, I compared a matrix of 126 rows (the +1 *de novo* frames) and 25 columns with a matrix of 126 rows (the ancestral frames) and 25 columns. LDA separated the +1 *de novo* frames from the ancestral ones with an accuracy of 97.2% (Fig. 6A). The error of classification was 2.4% for the *de novo* frames and 3.2% for the ancestral frames. The strong discriminant power of the linear function was confirmed with the validation test. When applied to 100 resamplings of the dataset, it yielded a mean accuracy of prediction of 95.8%.

Analysis of the other group (68 overlaps) revealed that the +2 *de novo* frames differ significantly from the ancestral ones for 23 composition features (13 with a significant enrichment and 10 with a significant depletion) (Supplementary Table S3). They were used as input data for LDA, which compared a matrix of 68 rows (+2 *de novo* frames) and 23 columns with a matrix of 68 rows (ancestral frames) and 23 columns. LDA separated the +2 *de novo* frames from the ancestral ones with an accuracy of 100% (Fig. 6B). This result was confirmed by the validation test, whose performance on 100 resamplings of the dataset was a mean accuracy of prediction of 98.0%.

3.9. Computational design of new variants of the apoptin from Chicken anemia virus (CAV) and of the X protein from Borna disease virus (BDV)

Search for homologs of the overlap capsid protein VP2/apoptin from CAV led to detection of Avian gyrovirus 2 (AGV2) and human Gyrovirus Tu789. These homologs are distantly related each other, as they show a mean nucleotide diversity of 39.2% and a mean amino acid diversity of 47.7% for VP2 and of 61.3% for apoptin. Despite the diversity, the apoptin of Gyrovirus Tu789 is able to trigger apoptosis in cancer cell lines at levels comparable to the CAV apoptin (Bullenkamp et al., 2012; Chaabane et al., 2017).

PLS-DA of the dataset (Fig. 2) assigned a score of -0.96 to the overlap from CAV, -0.76 to that from AGV2, and -0.94 to that from Gyrovirus Tu789 (PLS-DA score). LDA of the same dataset (Fig. 3) assigned a score of -42.0 to CAV, -40.2 to AGV2, and -41.2 to Gyrovirus Tu789 (LDA.1 score). LDA of ancestral and +1 *de novo* frames (Fig. 6A) assigned a score of 15.0 to the +1 frame encoding the CAV apoptin, 8.3 to that encoding the AGV2 apoptin, and 13.8 to that encoding the Gyrovirus Tu789 apoptin (LDA.2 score).

The algorithm randomly permuted the synonymous codons of the VP2 frame of CAV in 59 out of 121 codon positions (see Methods). Over a total of 150,000 rounds of permutation, the algorithm selected a total of 30,571 new overlaps with a +1 frame not interrupted by stop codons and meeting the following criteria: *i)* a PLS-DA score from -0.96 to -0.76; *ii)* a first LDA.1 score from -42.0 to -40.2; *iii)* a LDA.2 score from 8.3 to 15.0.

This high amount of new overlaps was notably reduced using a conservation criterion. Leaving out the 2 apoptosis-inducing domains, the multiple alignment of the apoptins from CAV, AGV2, and Gyrovirus Tu789 revealed conservation of 14 amino acids. Using as filter the presence of these amino acids, I obtained a set of 1111 apoptin variants with a mean amino acid identity of 87.5% (sd = 2.3%). The respective amino acid sequences are given in section A of Supplementary File S4.

A further reduction came from a more stringent selection of the 30,571 new overlaps. It made use of the closest PSA-DA and LDA scores assigned to CAV and Gyrovirus Tu 789: from -0.96 to -0.94 for PLS-DA score, from -42.0 to -41.2 for LDA.1 score, and from 13.8 to 15.0 for LDA.2 score. By this approach, the number of new overlaps was reduced to 1277 cases. Application of the conservation criterion yielded a total of 36 apoptin variants with a mean amino acid identity of 88.1% (sd = 2.2%). The corresponding amino acid sequences are given in section B of Supplementary File S4. The 7 most divergent apoptin variants, aligned to the CAV apoptin, are shown in Fig. 7.

Search for homologs of the overlap phosphoprotein/X protein from BDV led to detection of 6 homologs (from *Borna disease virus 2*, *Estrildid finch bornavirus 1*, *Aquatic bird bornavirus 1*, *Canary bornavirus 2*, *Variegated squirrel bornavirus 1*, and *Parrot bornavirus 5*). When compared each other, the 7 homologs showed a mean nucleotide diversity of 21.3% and a mean amino acid diversity of 16.9% for phosphoprotein and of 45.1% for X protein.

PLS-DA of the dataset (Fig. 2) assigned to the 7 overlaps a score ranging from -1.39 to -0.78 (PLS-DA score). LDA of the dataset (Fig. 3) assigned to them a score ranging from -48.1 to -42.5 (LDA.1 score). LDA of ancestral and +2 *de novo* frames (Fig. 6B) assigned to the seven +2 frames encoding X protein a score ranging from -59.7 to -47.6.

The algorithm randomly permuted the synonymous codons of the phosphoprotein frame of BDV in all its 70 codon positions. Over a total of 150,000 rounds of permutation, the algorithm selected a total of 4998 new overlaps with a +2 frame not interrupted by stop codons and meeting the following criteria: *i)* a PLS-DA score from -1.39 to -0.78; *ii)* a LDA.1 score from -48.1 to -42.5; *iii)* a LDA.2 score from -59.7 to -47.6. Using as additional filter a conservation criterion, that is the presence of the 16 conserved amino acids in the 7 homologous X proteins, I obtained a set of 6 variants with a mean amino acid identity of 65.0% (sd = 5.6%) (Fig. 8).

Szelechowski et al. (2014) found that a peptide covering the C-terminal end of the X protein from BDV (residues 59–87 in the sequence NC_001607, see Fig. 8) has a neuroprotective activity similar to that of the full-length protein, whereas no activity was found for peptides spanning residues 18–58. This finding suggested a further analysis. Using the conservation criterion, but limited to the region spanning residues 18–58, I selected a total of 30 shorter variants with a mean amino acid identity of 61.6% (sd = 6.4%). The corresponding amino acid sequences are given in Supplementary File S5.

4. Discussion

4.1. Creation of a dataset enriched in homologs allowed an exhaustive sequence analysis of viral overlapping genes

A limitation of our previous study (Pavesi et al., 2018) was that comparative analysis of overlapping and non-overlapping genes did not take into account the contribution to nucleotide diversity given by homologs, as it was carried out on 80 individual overlaps. Analysis of the present dataset (82 overlaps with the respective 237 homologs) led to detection in overlapping genes of a number of critical composition features that is about two-fold higher than that found previously (37 vs. 20). This finding made possible a PCA based on 37 variables. It confirmed that overlapping genes follow a common pattern of sequence composition (2 outliers out of 82 overlaps) more strongly than the previous PCA based on 20 variables (5 outliers out of 80 overlaps).

In addition, the present dataset made possible a more accurate evaluation of the pattern of symmetric/asymmetric evolution in overlapping genes. A previous study demonstrated that half of overlapping genes (32 out of 65) evolve in accordance to the asymmetric model (Pavesi, 2019). This analysis, however, underestimated the sequence diversity of overlapping genes, as it selected only one homolog per overlap and, moreover, with stringent criteria (an equal length in nucleotides and an alignment of the encoded proteins with virtually no gaps). The present analysis, which examined a number of homologs three-fold higher than that in the previous study (208 vs. 65), changed the trend in the direction of a prevalence of asymmetric (37 out of 65) vs. symmetric evolution (28 out of 65).

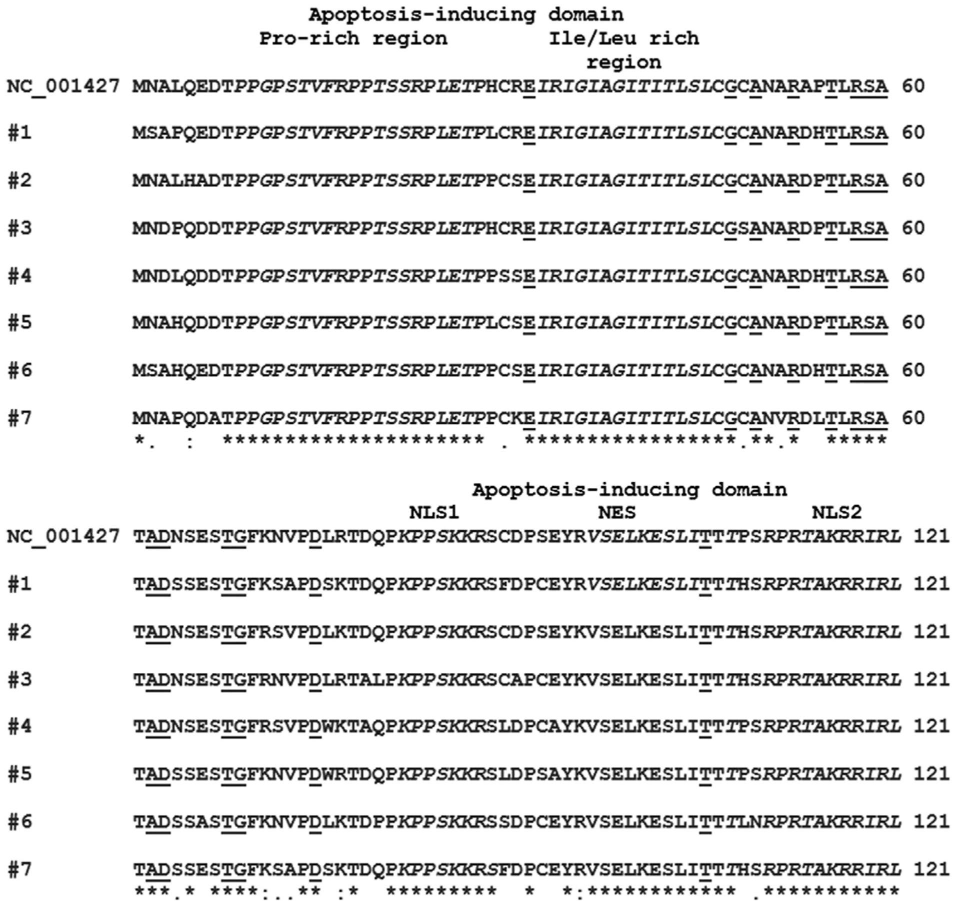
4.2. PLS-DA and LDA are a valuable tool for detecting new potential overlapping genes

The linear regression function given by PLS-DA, accounting for 23 composition features, and the linear function given by LDA, accounting for 21 composition features, separated overlapping from non-overlapping genes with an accuracy of 95.5% and 96.8%, respectively (Figs. 2 and 3). When considered jointly, the two methods correctly classified 94.2% of overlapping genes and 97.1% of non-overlapping genes (Fig. 4).

Thus, and as a complement to previous prediction methods (Firth, 2014; Sealfon et al., 2015; Schlub et al., 2018), the present model could be a powerful tool for detecting new potential overlapping genes in the viral genome sequences deposited in database. A joint application of PLS-DA and LDA to 8 overlapping genes previously classified as putative (Pavesi et al., 2018) revealed that 5 of them are *bona fide* overlapping genes (see Results). Application of the model to a set of overlapping genes predicted in viruses by Firth (2014) and Schlub et al. (2018) revealed that most of them (25 out of 33) are *bona fide* overlapping genes (data not shown). Finally, application of the model to a new potential overlapping gene found in *Hepatitis G virus* (Pavesi, 2000) confirmed that the NS5A region of the viral polyprotein is a dual-coding region, as it likely encodes a *de novo* gene product (data not shown).

Based on the finding that mammalian overlapping genes follows a composition bias similar to viral ones (Pavesi et al., 2018), the present model could also be useful for checking the wide amount of AltORFs detected in mammals by previous bioinformatics genome-wide studies (Chung et al., 2007; Ribrioux et al., 2008; Xu et al., 2010; Vanderperre et al., 2012).

Sequence analysis of about four thousand cancer exomes has revealed that the enrichment in synonymous nucleotide substitutions found in oncogenes depends on the fact that synonymous mutations frequently act as driver mutations in human cancers (Supek et al., 2014). By exploring the Supek's dataset, Brunet et al. (2018) suggested that the underlying cause of pathological synonymous mutations could



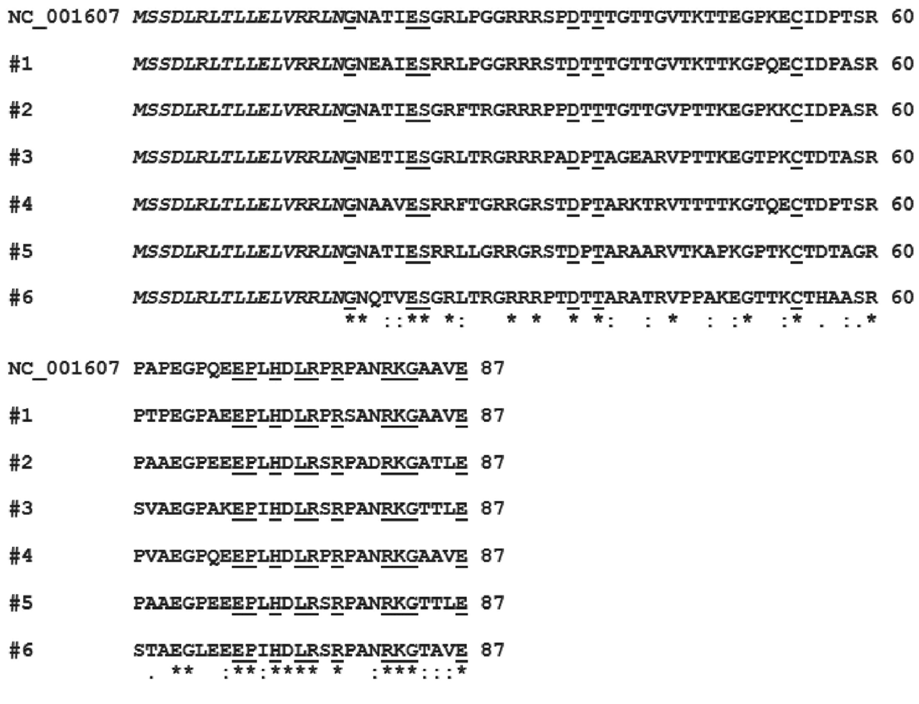
```
                    Apoptosis-inducing domain
                 Pro-rich region         Ile/Leu rich
                                            region
NC_001427 MNALQEDTPPGPSTVFRPPTSSRPLETPHCREIRIGIAGITITLSLCGCANARAPTLRSA 60

#1        MSAPQEDTPPGPSTVFRPPTSSRPLETPLCREIRIGIAGITITLSLCGCANARDHTLRSA 60

#2        MNALHADTPPGPSTVFRPPTSSRPLETPPCSEIRIGIAGITITLSLCGCANARDPTLRSA 60

#3        MNDPQDDTPPGPSTVFRPPTSSRPLETPHCREIRIGIAGITITLSLCGSANARDPTLRSA 60

#4        MNDLQDDTPPGPSTVFRPPTSSRPLETPPSSEIRIGIAGITITLSLCGCANARDHTLRSA 60

#5        MNAHQDDTPPGPSTVFRPPTSSRPLETPLCSEIRIGIAGITITLSLCGCANARDPTLRSA 60

#6        MSAHQEDTPPGPSTVFRPPTSSRPLETPPCSEIRIGIAGITITLSLCGCANARDHTLRSA 60

#7        MNAPQDATPPGPSTVFRPPTSSRPLETPPCKEIRIGIAGITITLSLCGCANVRDLTLRSA 60
          *.  :  *********************  .*****************.**.*  *****

                              Apoptosis-inducing domain
                        NLS1            NES              NLS2
NC_001427 TADNSESTGFKNVPDLRTDQPKPPSKKRSCDPSEYRVSELKESLITTTPSRPRTAKRRIRL 121

#1        TADSSESTGFKSAPDSKTDQPKPPSKKRSFDPCEYRVSELKESLITTTHSRPRTAKRRIRL 121

#2        TADNSESTGFRSVPDLKTDQPKPPSKKRSCDPSEYKVSELKESLITTTHSRPRTAKRRIRL 121

#3        TADNSESTGFRNVPDLRTALPKPPSKKRSCAPCEYKVSELKESLITTTHSRPRTAKRRIRL 121

#4        TADNSESTGFRSVPDWKTAQPKPPSKKRSLDPCAYKVSELKESLITTTPSRPRTAKRRIRL 121

#5        TADSSESTGFKNVPDWRTDQPKPPSKKRSLDPSAYKVSELKESLITTTPSRPRTAKRRIRL 121

#6        TADSSASTGFKNVPDLKTDPPKPPSKKRSSDPCEYRVSELKESLITTTLNRPRTAKRRIRL 121

#7        TADSSESTGFKSAPDSKTDQPKPPSKKRSFDPCEYRVSELKESLITTTHSRPRTAKRRIRL 121
          ***.* ****:..** :*  *********  *  *:************ .***********
```

Fig. 7. Alignment of the 7 most divergent apoptins yielded by simulation (from #1 to #7) with the apoptin of *Chicken anemia virus* (sequence NC_001427). Italic characters indicate the N-terminal apoptosis-inducing domain (Pro-rich region from residue 9 to 28 and Ile/Leu rich region from residue 33 to 46), the C-terminal apoptosis-inducing domain (NLS1 from residue 82 to 88, NES from residue 97 to 105, and NLS2 from residue 111 to 121), and a critical threonine site (residue 108). Due their critical role, these protein regions were excluded from random permutation. Underlined positions indicate the 14 conserved amino acids in the apoptins of CAV, AGV2, and Gyrovirus Tu789.

be an amino acid change in a "hidden" protein encoded by an alternative ORF in the same mRNA. Thus, another possible application of the model could be the search for "hidden" gene product in the Cancer Genome Atlas database.

4.3. Identification of 2 new potential overlapping genes in the genome of the new coronavirus SARS-CoV-2

Analysis of the genome sequence of SARS-CoV-2 using the scoring rules given by PLS-DA and LDA pointed out a few interesting features.



```
NC_001607 MSSDLRLTLLELVRRLNGNATIESGRLPGGRRRSPDTTTGTTGVTKTTEGPKECIDPTSR 60

#1        MSSDLRLTLLELVRRLNGNEAIESRRLPGGRRRSTDTTTGTTGVTKTTKGPQECIDPASR 60

#2        MSSDLRLTLLELVRRLNGNATIESGRFTRGRRRPPDTTTGTTGVPTTKEGPKKCIDPASR 60

#3        MSSDLRLTLLELVRRLNGNETIESGRLTRGRRRPADPTAGEARVPTTKEGTPKCTDTASR 60

#4        MSSDLRLTLLELVRRLNGNAAVESRRFTGRRGRSTDPTARKTRVTTTTKGTQECTDPTSR 60

#5        MSSDLRLTLLELVRRLNGNATIESRRLLGRRGRSTDPTARAARVTKAPKGPTKCTDTAGR 60

#6        MSSDLRLTLLELVRRLNGNQTVESGRLTRGRRRPTDTTARATRVPPAKEGTTKCTHAASR 60
                           ** ::** *:   * *  * *:  : *  : :*  :* . :.*

NC_001607 PAPEGPQEEPLHDLRPRPANRKGAAVE 87

#1        PTPEGPAEEPLHDLRPRSANRKGAAVE 87

#2        PAAEGPEEEPLHDLRSRPADRKGATLE 87

#3        SVAEGPAKEPIHDLRSRPANRKGTTLE 87

#4        PVAEGPQEEPLHDLRPRPANRKGAAVE 87

#5        PAAEGPEEEPLHDLRSRPANRKGTTLE 87

#6        STAEGLEEEPIHDLRSRPANRKGTAVE 87
           . **  :**:**** *  :***:::*
```

Fig. 8. Alignment of the 6 X protein selected variants (from #1 to #6) with the X protein from *Borna disease virus* (sequence NC_001607). Italic characters (residues 1–17) indicate the non-overlapping region of the X protein. Underlined positions indicate the 16 conserved amino acids in the overlapping region of the 7 homologous X proteins.

First at all, SARS-Cov-2 shows an overlap 3a protein/3b protein deeply different from the homologs from human and bat SARS-CoV. It may be that SARS-CoV-2 has lost the ability to encode a functional 3b protein. Indeed, the length of itts predicted 3b protein is about one quarter of that of human SARS-Cov, which acts as inhibitor of the host interferon response (Kopecky-Bromberg et al., 2007).

However, within the gene region that encodes 3a protein I identified a new potential overlapping gene, shifted of two nucleotide positions (Fig. 5B). Interestingly, this overlap is a feature unique to all the human SARS-CoV-2 genomes deposited in NCBI to date. If the hypothesis of a non-functional 3b protein in SARS-CoV-2 is correct, this finding suggests a replacement of the +1 frame encoding 3b with a +2 frame of unknown function. A similar mechanism has been described for the overlapping gene capsid protein VP2/apoptin in some human divergent gyroviruses (see Fig. 1 in Gia Phan et al., 2013).

In addition, within the gene region that encodes the nucleocapsid protein I identified a further new potential overlapping gene, shifted of one nucleotide position (Fig. 5A). Unlike the previous case, this putative overlap is a common feature of SARS-Cov-2 and SARS-CoV. The location of the two putative overlapping genes in the genome of SARS-CoV-2 suggests that the gene region encoding the 3a and nucleocapsid proteins could be a hotspot for overprinting. A similar feature was found previously in the "gene nursery" of deltaretroviruses (Pavesi et al., 2013).

4.4. Prediction of genealogy by codon usage is helpful to infer the relative age of overlapping genes

As shown in Supplementary Tables S1 and I could predict the genealogy for 46 out of 82 overlapping genes in the dataset. In addition to 9 overlaps whose genealogy was inferred only by phylogeny, Table S1 reports 37 overlaps in which prediction of ancestral and *de novo* frame came from codon-usage alone or combined with phylogeny. I would highlight two particular cases of gene overlap, as a paradigm of different evolutionary pathways.

The first is the overlap nucleocapsid protein/non-structural protein NSs from orthohantaviruses (6 homologs, see overlap #19 in Table S1). In all homologs, the codon-usage method demonstrated that the frame encoding NSs (the predicted *de novo* frame) has a codon bias significantly more distant to that of the viral genome than that of the frame encoding the nucleocapsid protein (P values ranging from 0.0001 to 0.005).

The other is the overlap envelope glycoprotein (EGP)/secreted glycoprotein (SGP) from ebolaviruses (6 homologs, see overlap#16 in Table S1). In 3 homologs, the codon-usage method demonstrated that the codon bias of the frame encoding SGP (the predicted *de novo* frame) is significantly more distant to that of the viral genome than the codon bias of the frame encoding EGP (P values ranging from 0.0001 to 0.025). In the remaining 3 homologs, the method yielded the same genealogy, but prediction was not supported significantly (P values ranging from 0.15 to 0.25).

Sabath et al. (2012) found that young *de novo* genes have a codon usage highly different from the rest of the genome and that older *de novo* genes tend to have a codon usage more adapted to that of the rest of the genome. Based on this finding, I can infer that the non-structural protein NSs of orthohantaviruses is encoded by a young *de novo* gene and that the secreted glycoprotein (SGP) of ebolaviruses by an older *de novo* gene. Table S1 reports several overlaps with a codon-usage statistics similar to that found in orthohantaviruses (e.g. overlaps #6 and #24) or to that found in ebolaviruses (e.g. overlaps #9 and #44). This information should be useful for inferring the relative age of other overlapping genes.

4.5. LDA should support the gene-novelty theory about the abundance of gene overlap in viruses

The finding that the length of viral genomes is negatively correlated with the length of the overlapping genes they contain could be a clue to support the gene-compression theory. This theory explains the abundance of gene overlap in viruses as a consequence of a physical constraint on genome length by the capsid (Chirico et al., 2010) or of a high mutation rate such that occurring in RNA viruses (Belshaw et al., 2007). As most mutations are deleterious, the high mutation rate will limit the genome size, and thus new genes must come from overprinting (Holmes, 2009).

However, the negative correlation I found between gene overlap and genome length, albeit significant, is weak (rho = -0.31). The strong correlation reported in previous studies (Belshaw et al., 2007; Chirico et al., 2010) could be artefactual, because it was a correlation between the length of the genome and the *ratio* of the length of overlaps to the length of the genome. As noted previously (Pavesi et al., 2018), using twice the same variable (the genome length) in a correlation test is statistically questionable, since the examined data are not independent.

In this study, LDA separated the ancestral overlapping frames from the *de novo* + 1 frames and from the +2 *de novo* frames with an accuracy close to 100% (Fig. 6). In detail, the *de novo* proteins encoded by the +1 frame are enriched in hydrophobic residues (leucine and methionine) and depleted in acidic residues (aspartic acid) (Supplementary Table S2). The *de novo* proteins encoded by the +2 frame are enriched in basic residues (arginine and histidine) and cysteine, and depleted in hydrophobic residues (leucine and methionine) (Supplementary Table S3).

These findings should support the gene-novelty theory, which states that the abundance of gene overlap in viruses is driven by selection pressures favouring the expression of new proteins with peculiar sequence properties (Rancurel et al., 2009; Brandes and Linial, 2016).

In the overlaps with known genealogy, the strong prevalence of the +1 *de novo* on the +2 *de novo* frames (36 and 11 cases, respectively, see last column in Table S1) can be explained by mutation bias. Indeed, analysis of the control set of non-overlapping genes (1,724,466 nt) revealed that start AUG codons are more frequent in the +1 frame (1 per 30 codons) than the +2 frame (1 per 110) and that stop codons are less frequent in the +1 frame (1 per 16 codons) than the +2 frame (1 per 12 codons). These results are in full accordance with those reported by Willis and Masel (2018).

Belshaw et al. (2007) proposed that prevalence of the +1 on the +2 frameshifts should result in a preponderance of NYR and YRN triplets, respectively. Indeed, we would expect to find more stop codons (TAA, TAG, and TGA) by chance in a YRN-rich (+2 frameshifted) sequence than in a NYR-rich (+1 frameshifted) sequence. In accordance to this proposal, analysis of the control set of non-overlapping genes (1,724,466 nt), shifted of 1 nucleotide position, showed a preponderance of the NYR triplet (31%) on the YNR triplet (26%). The same analysis, with a shift of 2 nucleotide positions, showed a preponderance of the YNR triplet (21%) on the NYR triplet (15%).

Finally, Belshaw et al. (2007) showed that in internal overlaps the +1 frameshifts are significantly more common than the +2 frameshifts (59 and 20 cases, respectively). This finding was fully confirmed by analysis of the 47 overlapping genes with known genealogy and of their homologs. Indeed, I found that the number of internal overlaps with a +1 *de novo* frame is almost 5-fold larger than that of internal overlaps with a +2 *de novo* frame (98 and 20 cases, respectively).

4.6. The new variants of apoptin and X protein are a useful benchmark for experimental studies

The selective anticancer toxicity of CAV-apoptin depends on a predominantly nuclear localization in tumor cells, whereas in normal cells it is detected mainly in cytoplasm (Danen-Van Oorschot et al., 2003). Apoptin has a N-terminal apoptosis-inducing domain, formed by a proline-rich segment (PRS) and a leucine-rich segment (LRS), and a C-terminal apoptosis-inducing domain, formed by a bi-partite nuclear localization sequence (NLS1 and NLS2) and a nuclear export sequence (NES).

Both the N- and C-terminal halves of apoptin can induce cell death on their own, albeit less strongly than the full-length protein (Danen-Van Oorschot et al., 2003). A truncated apoptin lacking residues 1–43 is a soluble, non-aggregating protein that maintains most of the properties of wild-type apoptin when transfected into human cancer cells (Rui-Martinez et al., 2017). Two other constructs, the first formed by PRS and NLSs and the other by LRS and NLSs, can induce selective apoptosis in a breast cancer cell line and in glioma cells, respectively (Shen Ni et al., 2013; Zhang et al., 2017).

The aim of these studies was to obtain deleted forms of CAV-apoptin retaining the selective anticancer activity of the full-length protein, yet with the advantage of an increased solubility (Rui-Fernandez et al., 2017) or a reduced immunogenicity (Zhang et al., 2017). However, all efforts in this direction have been conducted so far using the wild-type apoptin of CAV. By a computer simulation of the process of overprinting, the aim of the present study is to provide new variants of CAV-apoptin.

Using the scoring filters given by PLS-DA and LDA and a conservation criterion, I could reduce the huge amount of the new CAV-apoptins yielded by simulation to 1111 variants (section A of Supplementary File S4). Using a more stringent filter, I could obtain a set of 36 variants (section B of Supplementary File S4), 7 of them are shown in Fig. 7. They are a useful benchmark for testing the ability to induce selective cell death in transformed cells.

The X protein of BDV is encoded by a non-overlapping region (residues 1–17, see italic characters in Fig. 8) and an overlapping region (residues 18–87). The finding that its mitochondrial localization is mediated by residues 5–16 led to construction of two N-terminal manipulated mutants with an improved mitochondrial targeting and higher neuroprotective potential (Ferré et al., 2017). Szelechowski et al. (2014) found that a short peptide (residues 59–87) provides a protection against neurodegeneration in a mouse model of Parkinson's disease that is similar to the full-length X protein, yet with the advantage of a minimally invasive method of administration. The same study revealed that two other peptides (residues 1–29 and 30–59 respectively) did not exhibit any protection.

In addition to the scoring rules given by PLS-DA and LDA, selection of variants of the X protein of BDV was favoured by the finding that the overlap phosphoprotein/X protein undergoes strong asymmetric evolution. Indeed, the amino acid identity between the 7 homologous phosphoproteins (46 conserved residues) is three-fold greater than that between the 7 homologous X proteins (16 conserved residues).

Using the scoring filters given by PLS-DA and LDA and a conservation criterion, I could obtain a first set of 6 variants of the X-protein (Fig. 8). Using the same scoring filters, but limiting the conservation criterion to the region from residue 18 to 58, I could obtain a further set of 30 shorter variants (Supplementary File S5). Both datasets are a useful benchmark for experimental studies testing the neuroprotective potential, especially the second one due to the poorly invasive method of administration of short peptides *in vivo* (Szelechowski et al., 2014).

5. Conclusions

In the present study, a multivariate statistical analysis of a large dataset revealed new evolutionary features of viral overlapping genes. They show a common, and peculiar, pattern of nucleotide and amino acid composition. Thanks to discriminant analysis, overlapping genes can be separated from non-overlapping genes with high accuracy. Thus, the model I developed is a valuable tool for identifying new potential dual-coding regions in the viral genome sequences deposited in databases and, possibly, also in the eukaryotic genome sequences. A preliminary application of the model to SARS-CoV-2 led to prediction of 2 putative overlaps in the 3' genome region. The finding that the proteins encoded by the *de novo* frames differ significantly from those encoded by the ancestral frames supports the view that overprinting is a valuable source of genetic novelties. This view is corroborated by the notion that 2 *de novo* proteins (apoptin from CAV and X protein from BDV) can exert functions that are not virus-specific, such as a selective anticancer toxicity in human tumour cell lines and a protection against neurodegeneration in tissue culture, respectively. The search for variants of these proteins with an enhanced therapeutic effectiveness is an intriguing field of research. The contribution provided by this study could be helpful for future experimental studies. Finally, the wide collection of processed genome sequences given in Supplementary Files can be used by others as reference datasets for subsequent evolutionary studies (e.g. the relative age of overlapping genes or the occurrence of symmetric and asymmetric evolution in different regions of the same overlap).

Declaration of competing interest

None.

Acknowledgements

The author is grateful to Alessio Peracchi (University of Parma) and Alberto Vianelli (University of Insubria) for helpful suggestions. Special thanks to Dr. Gianmarco Del Vecchio for preparing the figures. The author thanks the anonymous referees for their helpful feedback and suggestions. This work has benefited from the equipment and framework of the COMP-HUB Initiative, funded by the 'Departments of Excellence' program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virol.2020.03.007.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Backendorf, C., Visser, A.E., de Boer, A.G., Zimmermann, R., Visser, M., Voskamp, P., Zhang, Y.H., Noteborn, M., 2008. Apoptin: therapeutic potential of an early sensor of carcinogenic transformation. Annu. Rev. Pharmacol. Toxicol. 48, 143–169.

Barrell, B.G., Air, G.M., Hutchison 3rd, C.A., 1976. Overlapping genes in bacteriophage phiX174. Nature 264, 34–41.

Belshaw, R., Pybus, O.G., Rambaut, A., 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 17, 1496–1504. https://doi:10.1101/gr.6305707.

Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., Motard, J., Roucou, X., 2013. An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. J. Biol. Chem. 288, 21824–21835. https://doi:10.1074/jbc.M113.472654.

Boehme, K.W., Hammer, K., Tollefson, W.C., Konopka-Anstadt, J.L., Kobayashi, T., Dermody, T.S., 2013. Nonstructural protein σ1s mediates reovirus-induced cell cycle arrest and apoptosis. J. Virol. 87, 12967–12979. https://doi:10.1128/JVI.02080-13.

Brandes, N., Linial, M., 2016. Gene overlapping and size constraints in the viral world. Biol. Direct 11, 26. https://doi:10.1186/s13062-016-0128-3.

Brereton, R.G., Lloyd, G.R., 2014. Partial least squares discriminant analysis: taking the magic away. Chemometrics 28, 213–225. https://doi:10.1002/cem.2609.

Brunet, M.A., Levesque, S.A., Hunting, D.J., Cohen, A.A., 2018. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. Genome Res. 28, 609–624. https://doi:10.1101/gr.230938.117.

Bullenkamp, J., Cole, D., Malik, F., Alkhatabi, H., Kulasekararaj, A., Odell, E.W., Farzaneh, F., Gäken, J., Tavassoli, M., 2012. Human Gyrovirus Apoptin shows a similar subcellular distribution pattern and apoptosis induction as the chicken anaemia virus derived VP3/Apoptin. Cell Death Dis. 3, e296. https://doi:10.1038/cddis.2012.34.

Carter, J.J., Daugherty, M.D., Qi, X., Bheda-Malge, A., Wipf, G.C., Robinson, K., Roman, A., Malik, H.S., Galloway, D.A., 2013. Identification of an overprinting gene in

Merkel cell polyomavirus provides evolutionary insight into the birth of viral gene. Proc. Natl. Acad. Sci. U.S.A. 110, 12744–12749. https://doi:10.1073/pnas.1303526110.

Castro, J., Ribo, M., Benito, A., Vilanova, M., 2018. Apoptin, a versatile protein with selective antitumor activity. Curr. Med. Chem. 25, 3540–3559. https://doi:10.2174/0929867325666180309112023.

Chaabane, W., Ghavami, S., Malecki, A., Los, M.J., 2017. Human gyrovirus-apoptin interferes with the cell cycle and induces G2/M arrest prior to apoptosis. Arch. Immunol. Ther. Exp. 65, 545–552. https://doi:10.1007/s00005-017-0464-8.

Chellappan, P., Vanitharani, R., Fauquet, C.M., 2005. MicroRNA-binding viral protein interferes with Arabidopsis development. Proc. Natl. Acad. Sci. U.S.A. 102, 10381–10386.

Chen, W., Calvo, P.A., Malide, D., Gibbs, J., Schubert, U., Bacik, I., Basta, S., O'Neill, R., Schickli, J., Palese, P., Henklein, P., Bennink, J.R., Yewdell, J.W., 2001. A novel influenza A virus mitochondrial protein that induces cell death. Nat. Med. 7, 1306–1312. https://doi:10.1038/nm1201-1306.

Chirico, N., Vianelli, A., Belshaw, R., 2010. Why genes overlap in viruses. Proc. Biol. Sci. 277, 3809–3817. https://doi:10.1098/rspb.2010.1052.

Chung, W.Y., Wadhawan, S., Szklarczyk, R., Pond, S.K., Nekrutenko, A., 2007. A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput. Biol. 3, e91. https://doi:10.1371/journal.pcbi.0030091.

Danen-Van Oorschot, A.A., Fischer, D.F., Grimbergen, J.M., Klein, B., Zhuang, S., Falkenburg, J.H., Backendorf, C., Quax, P.H., Van der Eb, A.J., Noteborn, M.H., 1997. Apoptin induces apoptosis in human transformed and malignant cells but not in normal cells. Proc. Natl. Acad. Sci. U.S.A. 94, 5843–5847.

Danen-Van Oorschot, A.A., Zhang, Y.H., Leliveld, S.R., Rohn, J.L., Seelen, M.C., Bolk, M.W., Van Zon, A., Erkeland, S.J., Abrahams, J.P., Mumberg, D., Noteborn, M.H., 2003. Importance of nuclear localization of apoptin for tumor-specific induction of apoptosis. J. Biol. Chem. 278, 27729–27736.

Delaye, L., Deluna, A., Lazcano, A., Becerra, A., 2008. The origin of a novel gene through overprinting in Escherichia coli. BMC Evol. Biol. 8, 31. https://doi:10.1186/1471-2148-8-31.

Fellner, L., Simon, S., Scherling, C., Witting, M., Schober, S., Polte, C., Schmitt-Kopplin, P., Keim, D.A., Scherer, S., Neuhaus, K., 2015. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. BMC Evol. Biol. 15, 283. https://doi:10.1186/s12862-015-0558-z.

Ferré, C.A., Davezac, N., Thouard, A., Peyrin, J.M., Belenguer, P., Miquel, M.C., Gonzalez-Dunia, D., Szelechowski, M., 2016. Manipulation of the N-terminal sequence of the Borna disease virus X protein improves its mitochondrial targeting and neuroprotective potential. Faseb. J. 30, 1523–1533. https://doi:10.1096/fj.15-279620.

Firth, A.E., 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. 42, 12425–12439. https://doi:10.1093/nar/gku981.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 8, 376–386. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gia Phan, T., Phung Vo, N., Sdiri-Loulizi, K., Aouni, M., Pothier, P., Ambert-Balay, K., Deng, X., Delwart, E., 2013. Divergent gyroviruses in the feces of Tunisian children. Virology 446, 346–348. https://doi:10.1007/s00705-015-2468-1.

Gibbs, A., Keese, P.K., 1995. Molecular Basis of Virus Evolution. Cambridge University Press, pp. 76–90.

Holmes, E.C., 2009. The Evolution and Emergence of RNA Viruses. Oxford University Press.

Hotelling, H., 1936. Relation between two sets of variates. Biometrika 38, 321–377. https://doi:10.1093/biomet/28.3-4.321.

Keese, P.K., Gibbs, A., 1992. Origin of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. U.S.A. 89, 9489–9493. https://doi.org/10.10173/pnas.89.20.9489.

Kopecky-Bromberg, S.A., Martínez-Sobrido, L., Frieman, M., Baric, R.A., Palese, P., 2007. Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists. J. Virol. 81, 548–557. https://doi:10.1128/JVI.01782-06.

Krakauer, D.C., 2000. Stability and evolution of overlapping genes. Evolution 54, 731–739.

Lachenbruch, P.A., Goldstein, M., 1979. Discriminant analysis. Biometrics 35, 69–85.

Lamb, R.A., 1980. Mapping of the two overlapping genes for polypeptides NS1 and NS2 on RNA segment 8 of influenza virus genome. Proc. Natl. Acad. Sci. U.S.A. 77, 1857–1861.

Lamb, R.A., Orvath, C.M., 1991. Diversity of coding strategies in influenza viruses. Trends Genet. 7, 261–266.

Lauber, C., Seitz, S., Mattei, S., Suh, A., Beck, J., Herstein, J., Börold, J., Salzburger, W., Kaderali, L., Briggs, J.A.G., Bartenschlager, R., 2017. Deciphering ther origin and evolution of Hepatitis B virus be means of a family of non-enveloped fish viruses. Cell Host Microbe 22 287-399. https://doi: 10.1016/j.chom.2017.07.019.

Lee, L.C., Liong, C.Y., Jemain, A.A., 2018. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategy and knowledge gaps. Analyst 143, 3526–3539. https://doi:10.1039/c8an00599k.

Leliveld, S.R., Zhang, Y.H., Rohn, J.L., Noteborn, M.H., Abrahams, J.P., 2003. Apoptin induces tumor-specific apoptosis as a globular multimer. J. Biol. Chem. 278, 9042–9051.

Li, F., Ding, S.W., 2006. Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu. Rev. Microbiol. 60, 503–531.

Mandic, M., Almunia, C., Vicel, S., Gillet, D., Janjic, B., Coval, K., Maillere, B., Kirkwood, J.M., Zarour, H.M., 2003. The alternative open reading frame of LAGE-1 gives rise to multiple promiscuous HLA-DR-restricted epitopes recognized by T-helper 1-type tumor-reactive CD4+ T cells. Canc. Res. 63, 6506–6515.

McFadden, N., Bailey, D., Carrara, G., Benson, A., Chaudhry, Y., Shortland, A., Heeney, J., Yarovinsky, F., Simmonds, P., Macdonald, A., Goodfellow, I., 2011. Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. PLoS Pathog. 7, e1002413. https://doi:10.1371/journal.ppat.1002413.

Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F., Baranov, P.V., 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 22, 2219–2229. https://doi:10.1101/gr.133249.111.

Miyata, T., Yasunaga, T., 1978. Evolution of overlapping genes. Nature 272, 532–535.

Morrison, D.F., 1976. Multivariate Statistical Methods. McGraw-Hill, New York.

Mouilleron, H., Delcourt, V., Roucou, X., 2016. Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. 44, 14–23. https://doi.org/10.1093/nar/gkv1218.

Normark, S., 1983. Overlapping genes. Annu. Rev. Genet. 17, 499–525.

Noteborn, M.H., Todd, D., Verschueren, C.A., de Gauw, H.W., Curran, W.L., Veldkamp, S., Douglas, A.J., McNulty, M.S., van der Eb, A.J., Koch, G., 1994. A single chicken anemia virus protein induces apoptosis. J. Virol. 68, 346–351.

Pavesi, A., De Iaco, B., Granero, M.I., Porati, A., 1997. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J. Mol. Evol. 44, 625–631.

Pavesi, A., 2000. Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. J. Mol. Evol. 50, 284–295.

Pavesi, A., Magiorkinis, G., Karlin, D.G., 2013. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of deltaretroviruses. PLoS Comput. Biol. 9, e1003162. https://doi:10.1371/journal.pcbi.1003162.

Pavesi, A., 2015. Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution. J. Gen. Virol. 96, 3577–3586.

Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., Karlin, D., 2018. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. PloS One 13, e0202513. https://doi:10.1371/journal.pone.0202513.

Pavesi, A., 2019. Asymmetric evolution in viral overlapping genes is a source of selective protein adaptation. Virology 532, 39–47. https://doi:10.1016/j.virol.2019.03.017.

Peleg, O., Kirzhner, V., Trifonov, E., Bolshoy, A., 2004. Overlapping messages and survivability. J. Mol. Evol. 59, 520–527.

Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., Karlin, D., 2009. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J. Virol. 83, 10719–10736. https://doi:10.1128/JVI.0059-09.

Ribrioux, S., Brungger, A., Baumgarten, B., Seuwen, K., John, M.R., 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genom. 9, 122. https://doi:10.1186/1471-2164-9-122.

Ringnér, M., 2008. What is principal component analysis? Nat. Biotechnol. 26, 303–304. https://doi: 10.1038/nbt0308-303.

Rohn, J.L., Zhang, Y.H., Aalbers, R.I., Otto, N., Den Hertog, J., Henriquez, N.V., Van De Velde, C.J., Kuppen, P.J., Mumberg, D., Donner, P., Noteborn, M.H., 2002. A tumor-specific kinase activity regulates the viral death protein apoptin. J. Biol. Chem. 277, 50820–50827.

Rosenberg, S.A., Tong-On, P., Li, Y., Riley, J.P., El-Gamil, L., Parkhurst, M.R., Robbins, P.F., 2002. Identification of BING-4 cancer antigen from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. J. Immunol. 168, 2402–2407.

Rosner, B., 1975. On the detection of many outliers, 1975. Technometrics 17, 221–227 1975.

Ruiz-Martínez, S., Castro, J., Vilanova, M., Bruix, M., Laurents, D.V., Ribó, M., Benito, A., 2017. A truncated apoptin protein variant selectively kills cancer cells. Invest. N. Drugs 35, 260–268.

Sabath, N., Wagner, A., Karlin, D., 2012. Evolution of viral proteins originated de novo by overprinting. Mol. Biol. Evol. 29, 3767–3780. https://doi:10.1093/molbev/mss179.

Schlub, T.E., Buchmann, J.P., Holmes, E.C., 2018. A simple method to detect candidate overlapping genes in viruses using single genome sequences. Mol. Biol. Evol. 35, 2572–2581. https://doi: 10.1093/molbev/msy155.

Sealfon, R.S., Lin, M.F., Jungreis, I., Wolf, M.Y., Kellis, M., Sabeti, P.C., 2015. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 16, 38. https://doi:10.1186/s13059-015-0603-7.

Shen Ni, L., Allaudin, Z.N., Mohd Lila, M.A., Othman, A.M., Othman, F.B., 2013. Selective apoptosis induction in MCF-7 cell line by truncated minimal functional region of Apoptin. BMC Canc. 13, 488. https://doi: 10.1186/1471-2407-13-488.

Siegel, S., Castellan Jr., N.J., 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Inc., New-York, pp. 87–95.

Sievers, F., Higgins, D.G., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079, 105–116. https://doi: 10.1007/978-1-62703-646-7_6.

Slager, E.H., Borghi, M., van der Minne, C.E., Aarnoudse, C.A., Havenga, M.J.E., Schrier, P.I., Osanto, S., Griffioen, M., 2003. CD4+ Th2 cell recognition of HLA-DR-restricted epitopes derived from CAMEL: a tumor antigen translated in an alternative open reading frame. J. Immunol. 170, 1490–1497.

Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G., Serody, J.S., 2019. Alternative tumour-specific antigens. Nat. Rev. Canc. 8, 465–478. https://doi:10.1038/s41568-019-0162-4.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., Lehner, B., 2014. Synonymous mutations frequently act as driver mutations in human cancers. Cell 156, 1324–1335. https://doi:10.1016/j.cell.2014.01.051.

Szelechowski, M., Bétourné, A., Monnet, Y., Ferré, C.A., Thouard, A., Foret, C., Peyrin, J.M., Hunot, S., Gonzalez-Dunia, D., 2014. A viral peptide that targets mitochondria protects against neuronal degeneration in models of Parkinson's disease. Nat.

Commun. 5, 5181. https://doi:10.1038/ncomms6181.

Taliansky, M., Roberts, I.M., Kalinina, N., Ryabov, E.V., Raj, S.K., Robinson, D.J., Oparka, K.J., 2003. An umbraviral protein, involved in long-distance RNA movement, binds viral RNA and forms unique, protective ribonucleoprotein complexes. J. Virol. 77, 3031–3040.

Vanderperre, B., Lucier, J.F., Roucou, X., 2012. HAltORF : a Database of Predicted Out-Of-Frame Alternative Open Reading Frames in Human. Database, 2012, bas025. https://doi: 10.1093/database/bas025.

Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M., Roucou, X., 2013. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PloS One 8, e70698. https://doi:10.1371/journal.pone.0070698.

van Knippenberg, I., Carlton-Smith, C., Elliott, R.M., 2010. The N-terminus of Bunyamwera orthobunyavirus NSs protein is essential for interferon antagonism. J. Gen. Virol. 91, 2002–2006. https://doi:10.1099/vir.0.021774-0.

Vargason, J.M., Szittya, G., Burgyan, J., Hall, T.M., 2003. Size selective recognition of siRNA by an RNA silencing suppressor. Cell 115, 799–811.

Wang, R.F., Parkhurst, M.R., Kawakami, Y., Robbins, P.F., Rosenberg, S.A., 1996. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. J. Exp. Med. 183, 1131–1140.

Wang, R.F., Johnson, S.L., Zeng, G., Topalian, S.L., Schwartzentruber, D.J., Rosenberg, S.A., 1998. A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames. J. Immunol. 161, 3598–3606.

Wensman, J.J., Munir, M., Thaduri, S., Hörnaeus, K., Rizwan, M., Blomström, A.L., Briese, T., Lipkin, W.I., Berg, M., 2013. The X protein of bornaviruses interfere with type I interferon signalling. J. Gen. Virol. 94, 263–269. https://doi: 10.1099/vir.0.047175-0.

Willis, S., Masel, J., 2018. Gene birth contributes to structural disorder encoded by overlapping genes. Genetics 210, 303–313. https://doi: 10.1534/genetics.118.301249.

Xu, H., Wang, P., Fu, Y., Zheng, Y., Tang, Q., Si, L., You, J., Zhang, Z., Zhu, Y., Zhou, L., Wei, Z., Lin, B., Hu, L., Kong, X., 2010. Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res. 20, 445–457. https://doi: 10.1038/cr.2010.25.

Zhang, L., Zhao, H., Cui, Z., Lv, Y., Zhang, W., Ma, X., Zhang, J., Sun, B., Zhou, D., Yuan, L., 2017. A peptide derived from apoptin inhibits glioma growth. Oncotarget 8, 31119–31132. htpps://doi:10.18632/oncotarget.16094.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. https://doi:10.1038/s41586-020-2012-7.